Adaptive Biomarker Driven Clinical Trials

Richard Simon, D.Sc.
Chief, Biometric Research Branch
National Cancer Institute
http://brb.nci.nih.gov

Adaptive Clinical Trials

- Sample size re-estimation
- Randomization response adaptive
- Stratification adaptive
- Treatment group adaptive
- Target population adaptive

VOLUME 29 · NUMBER 6 · FEBRUARY 20 2011

JOURNAL OF CLINICAL ONCOLOGY

STATISTICS IN ONCOLOGY

Outcome-Adaptive Randomization: Is It Useful?

Edward L. Korn and Boris Freidlin

See accompanying editorial on page 606

From the National Cancer Institute, Bethesda, MD.

Submitted June 18, 2010; accepted August 31, 2010; published online ahead of print at www.jco.org on December 20, 2010.

Authors' disclosures of potential conflicts of interest and author contributions are found at the end of this article.

Corresponding author: Edward L. Korn, PhD, Biometric Research Branch, EPN-8129, National Cancer Institute, Bethesda, MD 20892; e-mail: korne@ctep.nci.nih.gov.

Published by the American Society of Clinical Oncology

0732-183X/11/2906-771/$20.00

DOI: 10.1200/JCO.2010.31.1423

ABSTRACT

Outcome-adaptive randomization is one of the possible elements of an adaptive trial design in which the ratio of patients randomly assigned to the experimental treatment arm versus the control treatment arm changes from 1:1 over time to randomly assigning a higher proportion of patients to the arm that is doing better. Outcome-adaptive randomization has intuitive appeal in that, on average, a higher proportion of patients will be treated on the better treatment arm (if there is one). In both the randomized phase II and phase III settings with a short-term binary outcome, we compare outcome-adaptive randomization with designs that use 1:1 and 2:1 fixed-ratio randomizations (in the latter, twice as many patients are randomly assigned to the experimental treatment arm). The comparisons are done in terms of required sample sizes, the numbers and proportions of patients having an inferior outcome, and we restrict attention to the situation in which one treatment arm is a control treatment (rather than the less common situation of two experimental treatments without a control treatment). With no differential patient accrual rates because of the trial design, we find no benefits to outcome-adaptive randomization over 1:1 randomization, and we recommend the latter. If it is thought that the patient accrual rates will be substantially higher because of the possibility of a higher proportion of patients being randomly assigned to the experimental treatment (because the trial will be more attractive to patients and clinicians), we recommend using a fixed 2:1 randomization instead of an outcome-adaptive randomization.

J Clin Oncol 29:771-776. Published by the American Society of Clinical Oncology

Table 2. Average Proportion of Responders, No. of Nonresponders, and Overall Proportion Treated on the Experimental Arm for Various Randomized Phase II Trial Designs, Some of Which Use Adaptive Randomization

		Fixed Sample Size				Adaptive Randomization (N = 140)					
						Capped at 80% Assignment Probability			Capped at 90% Assignment Probability		
Response Rates by Arm		1:1 (n = 132)		2:1 (n = 153)							
Control Arm	Experimental Arm	P (responders) %	No. of Nonresponders	P (responders) %	No. of Nonresponders	P (responders) %	No. of Nonresponders	Overall % Treated on Experimental Arm	P (responders) %	No. of Nonresponders	Overall % Treated on Experimental Arm
0.2	0.2	20.0	105.6	20.0	122.4	20.0	112.0	50.0	20.0	112.0	50.0
0.2	0.3	25.0	99.0	26.6	112.2	26.0	103.6	59.7	26.0	103.6	60.3
0.2	**0.4**	**30.0**	**92.4**	**33.3**	**102.0**	**33.2**	**93.5**	**66.2**	**33.7**	**92.9**	**68.2**
0.2	0.5	35.0	85.8	40.0	91.8	41.0	82.6	69.9	42.1	81.1	73.6

NOTE. Adaptive randomization uses the method of Thall and Wathen[12] but with no early stopping. One-sided type 1 error = 10%, power = 90% at 20% *v* 40% response rates; results based on 500,000 simulations. Characteristics of trial designs corresponding to the trial alternative hypothesis are in bold type. P (responders) % is the average proportions of responders given as a percentage.

Table 4. Average Proportion of Responders and No. of Nonresponders for Various Randomized Phase III Trial Designs

1-Year Survival Rates		Fixed Sample Size (1:1) (n = 522; 261:261)		Fixed Sample Size (2:1) (n = 573; 382:191)		Adaptive Randomization: Block-Stratified Analysis and Randomization Capped at 80% Assignment Probability (n = 748; block size = 50)		
Control Arm	Experimental Arm	P (responders) %	No. of Nonresponders	P (responders) %	No. of Nonresponders	P (responders) %	No. of Nonresponders	Overall % Treated on Experimental Arm
0.8	0.8	80.0	104.4	80.0	114.6	80.0	149.6	50.0
0.8	0.85	82.5	91.4	83.3	95.5	83.3	124.6	66.9
0.8	**0.9**	**85.0**	**78.3**	**86.7**	**76.4**	**87.5**	**93.4**	**75.1**

NOTE. Adaptive randomization uses the method of Thall and Wathen[12] but with no early stopping. One-sided type 1 error = 2.5%, power = 90% at 80% *v* 90% 1-year survival rates; results based on 500,000 simulations. Characteristics of trial designs corresponding to the trial alternative hypothesis are in bold type. P (responders) % is the average proportions of responders given as a percentage.

Statistics and Probability Letters 81 (2011) 767–772

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Using randomization tests to preserve type I error with response adaptive and covariate adaptive randomization

Richard Simon [a,*], Noah Robin Simon [b]

[a] Biometric Research Branch, National Cancer Institute, 9000 Rockville Pike, Bethesda, MD 20892-7434, United States
[b] Department of Statistics, Stanford University, Stanford, CA 94305, United States

ARTICLE INFO

Article history:
Available online 5 January 2011

Keywords:
Response adaptive randomization
Adaptive stratification
Clinical trials

ABSTRACT

We demonstrate that clinical trials using response adaptive randomized treatment assignment rules are subject to substantial bias if there are time trends in unknown prognostic factors and standard methods of analysis are used. We develop a general class of randomization tests based on generating the null distribution of a general test statistic by repeating the adaptive randomized treatment assignment rule holding fixed the sequence of outcome values and covariate vectors actually observed in the trial. We develop broad conditions on the adaptive randomization method and the stochastic mechanism by which outcomes and covariate vectors are sampled that ensure that the type I error is controlled at the level of the randomization test. These conditions ensure that the use of the randomization test protects the type I error against time trends that are independent of the treatment assignments. Under some conditions in which the prognosis of future patients is determined by knowledge of the current randomization weights, the type I error is not strictly protected. We show that response adaptive randomization can result in substantial reduction in statistical power when the type I error is preserved. Our results also ensure that type I error is controlled at the level of the randomization test for adaptive stratification designs used for balancing covariates.

Published by Elsevier B.V.

Table 1
Type I error for Mann–Whitney test.

	No time trend	Time trend
Simple randomization	0.046	0.050
Adaptive randomization	0.049	0.205

Table 1 shows the result of a simple simulation of two arm clinical trials with data for $n = 50$ patients. The test used for comparing the groups is based on the Mann–Whitney statistic. The observed difference was considered statistically significant if the large sample normal approximation was significant at a one-sided 5% level. For the first row, treatment assignment was based on simple equally weighted non-adaptive randomization. For the first column of the table, the outcomes $y_1, y_2, \ldots, y_n$ are independent and normally distributed with mean zero and variance 1; there are no measured covariates and no treatment effect. In this case, the type I error, estimated from 10,000 replicated trials, approximates the nominal 5% significance level used for the tests. The last column shows results when there is an unknown time trend. That is, y_i was normally distributed with mean $10i/n$ and variance 1. Again there were no measured covariates and no treatment effect. With or without time trend, using equally weighted non-adaptive randomization, the proportion of the 10,000 simulation replications in which the null hypothesis was rejected is approximately 0.05, and the small discrepancy is within the limitations of the number of replications and the accuracy of the large sample approximation to the Mann–Whitney statistic for clinical trials of only 50 patients.

The second row of Table 1 shows results for similar trials using a response adaptive randomization method. We assume that there is no delay in observing responses so H_i consists of outcomes and treatment assignments for patients $1, 2, \ldots, i-1$. The first 10 patients are assigned treatment using simple equally weighted randomization. For subsequent patients the randomization weight $g(H_i)$ is the standardized Mann–Whitney statistic for comparing outcomes for the two treatments using data for patients $1, 2, \ldots, i-1$. This standardized statistic equals the sum of the ranks for outcomes on treatment $c = 1$ minus $n_1(n_1+1)$ divided by $n_1 n_0$ where n_1 and n_0 denote the number of the first $i-1$ patients who received treatments 1 and 0 respectively. This standardized statistic takes values in the range 0–1.

3. A significance test based on the randomization distribution

Let $dF_{\underline{c}|\underline{z}}$ denote the distribution of the sequence of treatment assignments $\underline{c} = (c_1, \ldots, c_n)$ conditional on $\underline{z} = ((x_1, y_1), \ldots, (x_n, y_n))$, the sequence of covariate vectors $\underline{x} = (x_1, \ldots, x_n)$ and outcomes $\underline{y} = (y_1, \ldots, y_n)$. One can sample from $dF_{\underline{c}|\underline{z}}$ under the null hypothesis by holding fixed the sequence of covariate vectors and outcomes for the patients in the clinical trial and re-randomizing all of the patients using the probabilistic treatment assignment mechanism determined by the adaptive algorithm. The sequence of treatment assignments sampled will in general depend on the sequence of covariate vectors and outcomes and these are kept fixed.

Let $dF_{T(\underline{z})}$ denote the distribution of the test statistic T induced when the vector of treatment assignments is drawn from $dF_{\underline{c}|\underline{z}}$. This induced distribution can be used as a null distribution for the test statistic computed from the data using the treatment assignments actually used in the clinical trial. For a one-sided significance test of level α of the null hypothesis against the alternative that treatment 1 is superior, we use as critical value for the test statistic the $100(1-\alpha)$th percentile of $dF_{T(\underline{z})}$, i.e. $F^{-1}_{T(\underline{z})}(1-\alpha)$.

Table 2
Type I error for adaptive randomization test.

	No time trend	Time trend
Adaptive randomization	0.049	0.050

Theorem 1. *Let* $\underline{z} = ((x_1, y_1), \ldots, (x_n, y_n))$ *be a sequence of pairs of covariate vectors and outcomes and let* $\mathrm{d}F_{\underline{z},\underline{c}}$ *denote the joint distribution of* $\underline{z}$ *and the vector of treatment assignments* $\underline{c}$*. Let* $T(\underline{z}, \underline{c})$ *denote the value of the test statistic computed on the data and* $F_{T(\underline{z})}$ *denote the distribution function of the null distribution of the test statistic* T *induced by the randomization process conditional on* $\underline{z}$*. For each* $i \in \{1, \ldots, n\}$*, we assume that conditional on* $((x_1, y_1), \ldots, (x_{i-1}, y_{i-1}))$*,* (x_i, y_i) *is independent of* $(c_1, \ldots, c_{i-1})$*. Then under the null hypothesis,*

$$\Pr_{\underline{z},\underline{c}}\left[T(\underline{z}, \underline{c}) \leq F^{-1}_{T(\underline{z})}(1-\alpha)\right] \geq 1 - \alpha. \tag{1}$$

Table 3
Power for adaptive randomization.

Total sample size (n)	Time trend (β)	Power cap = 1	Power cap = 0.67
50	0	0.76	0.82
50	1	0.73	0.80
100	0	0.85	0.85
100	1	0.83	0.84

BIOMETRICS *31*, 103–115
March 1975

SEQUENTIAL TREATMENT ASSIGNMENT WITH BALANCING FOR PROGNOSTIC FACTORS IN THE CONTROLLED CLINICAL TRIAL

STUART J. POCOCK

Statistical Laboratory, SUNY at Buffalo, Amherst, New York 14226, U.S.A.

RICHARD SIMON

National Cancer Institute, Bethesda, Maryland 20014

SUMMARY

In controlled clinical trials there are usually several prognostic factors known or thought to influence the patient's ability to respond to treatment. Therefore, the method of sequential treatment assignment needs to be designed so that treatment balance is simultaneously achieved across all such patient factors. Traditional methods of restricted randomization such as "permuted blocks within strata" prove inadequate once the number of strata, or combinations of factor levels, approaches the sample size. A new general procedure for treatment assignment is described which concentrates on minimizing imbalance in the distributions of treatment numbers within the levels of each individual prognostic factor. The improved treatment balance obtained by this approach is explored using simulation for a simple model of a clinical trial. Further discussion centers on the selection, predictability and practicability of such a procedure.

Adaptive Treatment Selection

BIOMETRICS 45, 537–547
June 1989

A Two-Stage Design for Choosing Among Several Experimental Treatments and a Control in Clinical Trials

Peter F. Thall

Statistics/Computer & Information Systems Department,
George Washington University, Washington, D.C. 20052, U.S.A.

Richard Simon

Biometric Research Branch, Cancer Therapy Evaluation Program, DCT, NCI, NIH,
6130 Executive Boulevard, Rockville, Maryland 20892, U.S.A.

and

Susan S. Ellenberg

Biostatistics Research Branch, AIDS Program, NIAID, NIH,
6003 Executive Boulevard, Rockville, Maryland 20892, U.S.A.

SUMMARY

In clinical trials where several experimental treatments are of interest, the goal may be viewed as identification of the best of these and comparison of that treatment to a standard control therapy. However, it is undesirable to commit patients to a large-scale comparative trial of a new regimen without evidence that its therapeutic success rate is acceptably high. We propose a two-stage design in which patients are first randomized among the experimental treatments, and the single treatment having the highest observed success rate is identified. If this highest rate falls below a fixed cutoff then the trial is terminated. Otherwise, the "best" new treatment is compared to the control at a second stage. Locally optimal values of the cutoff and the stage-1 and stage-2 sample sizes are derived by minimizing expected total sample size. The design has both high power and high probability of terminating early when no experimental treatment is superior to the control. Numerical results for implementing the design are presented, and comparison to Dunnett's (1984, in *Design of Experiments: Ranking and Selection*, T. J. Santner and A. C. Tamhane (eds), 47–66; New York: Marcel Dekker) optimal one-stage procedure is made.

Table 1

Designs for fixed power β and size $\alpha = .05$, $\delta_1 = .05$, $\delta_2 = .20$

θ_0^*	K	n_1	n_2	λ	$E_0(N)$	$E(N)$	N_{max}	$1 - \pi_0$	β_1	β_2	β
.2	2	28	89	.290–.320	86.6	150.3	234	.8280	.8041	.8704	.70
		32	97	.285–.310	97.0	168.4	258	.8284	.8381	.8949	.75
		40	99	.280–.300	113.0	190.0	278	.8326	.8869	.9021	.80
.2	3	31	98	.295–.320	133.5	202.4	289	.7925	.7785	.8992	.70
		38	102	.290–.315	149.7	227.0	318	.8245	.8231	.9112	.75
		45	111	.290–.310	167.9	256.6	357	.8518	.8580	.9324	.80
.2	4	34	105	.295–.320	183.1	256.6	346	.7740	.7631	.9173	.70
		41	109	.295–.315	205.9	288.0	382	.8073	.8082	.9280	.75
		48	120	.295–.310	231.1	325.9	432	.8362	.8441	.9478	.80
.4	2	33	98	.485–.515	110.5	177.5	262	.7725	.8250	.8485	.70
		39	108	.490–.510	119.8	198.6	294	.8063	.8515	.8808	.75
		47	119	.490–.510	131.3	224.1	332	.8430	.8800	.9091	.80
.4	3	38	111	.505–.525	162.0	239.3	336	.7828	.7876	.8888	.70
		44	123	.505–.520	177.9	268.3	378	.8123	.8182	.9166	.75
		51	129	.495–.505	205.7	302.5	411	.7952	.8612	.9290	.80
.4	4	40	120	.505–.525	223.7	304.1	400	.7341	.7681	.9113	.70
		48	128	.505–.520	248.0	340.3	448	.7795	.8099	.9260	.75
		56	142	.505–.515	275.6	384.5	508	.8172	.8437	.9482	.80
.6	2	26	86	.695–.730	91.3	149.3	224	.7717	.8148	.8591	.70
		32	90	.690–.715	102.8	166.6	244	.7822	.8612	.8709	.75
		39	99	.695–.715	111.1	187.3	276	.8313	.8875	.9014	.80
.6	3	30	94	.705–.730	138.1	200.6	278	.7436	.7885	.8878	.70
		37	100	.705–.725	151.1	223.9	311	.7987	.8288	.9049	.75
		43	107	.700–.720	170.6	251.7	343	.8055	.8661	.9237	.80
.6	4	34	99	.710–.735	187.6	253.7	334	.7379	.7763	.9017	.70
		41	105	.710–.730	207.3	283.8	374	.7919	.8178	.9171	.75
		47	113	.705–.720	233.7	319.2	414	.7976	.8545	.9362	.80

Table 3

Comparison of required sample sizes for Dunnett's single-stage and proposed two-stage designs at $\alpha = .05$, $\theta_0 = .20$, $\delta_1 = .05$, $\delta_2 = .20$

		Dunnett	Two-stage		
	β	N_t	$E_0(N)$	$E(N)$	N_{max}
$K = 2$	.70	189	86.6	150.3	234
	.75	213	97.0	168.4	258
	.80	240	113.0	190.0	278
$K = 3$	.70	284	133.5	202.4	289
	.75	316	149.7	227.0	318
	.80	356	167.9	256.6	357
$K = 4$	.70	385	183.1	256.6	346
	.75	425	205.9	288.0	382
	.80	475	231.1	325.9	432

Biometrika (1988), **75**, 2, *pp.* 303–10
Printed in Great Britain

Two-stage selection and testing designs for comparative clinical trials

BY PETER F. THALL
Department of Statistics/Computer & Information Systems, George Washington University, Washington, D.C. 20892, U.S.A.

RICHARD SIMON AND SUSAN S. ELLENBERG
National Cancer Institute, Bethesda, Maryland 20892, U.S.A.

SUMMARY

A two-stage design which selects the best of several experimental treatments and compares it to a standard control is proposed. The design allows early termination with acceptance of the global null hypothesis. Optimal sample size and cut-off parameters are obtained by minimizing expected total sample size for fixed significance level and power.

Table 1. *Designs having minimal* $E(N)$ *for given* K, θ_0, $1-\beta^*$, $\alpha = 0{\cdot}05$, $\delta_1 = 0{\cdot}05$ *and* $\delta_2 = 0{\cdot}20$

K	θ_0	$1-\beta^*$	n_1	n_2	y_1	y_2	$E(N)$	N_{max}	τ_0	γ^*
2	0·2	0·70	30	44	0·787	1·787	141·71	178	0·667	0·030
		0·75	36	44	0·730	1·818	163·71	196	0·640	0·026
		0·80	40	52	0·689	1·812	187·64	224	0·626	0·025
2	0·4	0·70	36	50	0·684	1·811	172·99	208	0·588	0·027
		0·75	40	58	0·590	1·808	197·36	236	0·575	0·026
		0·80	47	62	0·580	1·822	226·53	265	0·557	0·023
2	0·6	0·70	27	45	0·578	1·794	139·62	181	0·589	0·028
		0·75	31	50	0·543	1·798	159·54	193	0·584	0·026
		0·80	36	55	0·500	1·803	183·68	218	0·563	0·023
3	0·2	0·70	33	55	0·762	1·902	205·09	242	0·571	0·046
		0·75	38	59	0·709	1·916	233·33	270	0·548	0·042
		0·80	48	57	0·835	1·926	266·97	306	0·619	0·035
3	0·4	0·70	39	64	0·591	1·917	247·09	284	0·492	0·042
		0·75	47	63	0·550	1·944	280·89	314	0·469	0·036
		0·80	52	75	0·500	1·936	320·37	358	0·457	0·034
3	0·6	0·70	32	51	0·530	1·928	201·04	230	0·496	0·039
		0·75	37	55	0·509	1·935	227·94	258	0·490	0·036
		0·80	42	62	0·472	1·938	260·28	292	0·471	0·032
4	0·2	0·70	36	61	0·721	1·984	267·26	302	0·496	0·055
		0·75	44	62	0·868	1·983	303·14	344	0·583	0·049
		0·80	51	65	0·800	2·004	345·64	385	0·555	0·043
4	0·4	0·70	45	69	0·675	1·987	321·58	363	0·512	0·049
		0·75	49	77	0·550	2·004	364·32	399	0·404	0·046
		0·80	58	84	0·700	2·001	414·47	458	0·471	0·041
4	0·6	0·70	35	58	0·529	2·002	262·05	291	0·440	0·047
		0·75	42	61	0·717	2·000	296·28	332	0·520	0·042
		0·80	48	66	0·655	2·014	337·10	372	0·486	0·037

$E(N) = \frac{1}{2}\{E(N|H_0) + E(N|\theta^*)\}$; $N_{max} = (K+1)n_1 + 2n_2$; $\tau_0 = \text{pr}(T_1 \leq y_1 | H_0)$
$\gamma^* = \text{pr}$ (choose suboptimal $E_r | \theta^*$)

Target Population Adaptive

Standard Paradigm of Phase III Clinical Trials

- Broad eligibility
- Base primary analysis on ITT eligible population
- Don't size for subset analysis, allocate alpha for subset analysis or trust subset analysis
  - Only do subset analysis if overall treatment effect is significant and interaction is significant

Standard Paradigm Sometimes Leads to

- Large NNT
- Small average treatment effects
- Inconsistency in results among studies
- False negative studies

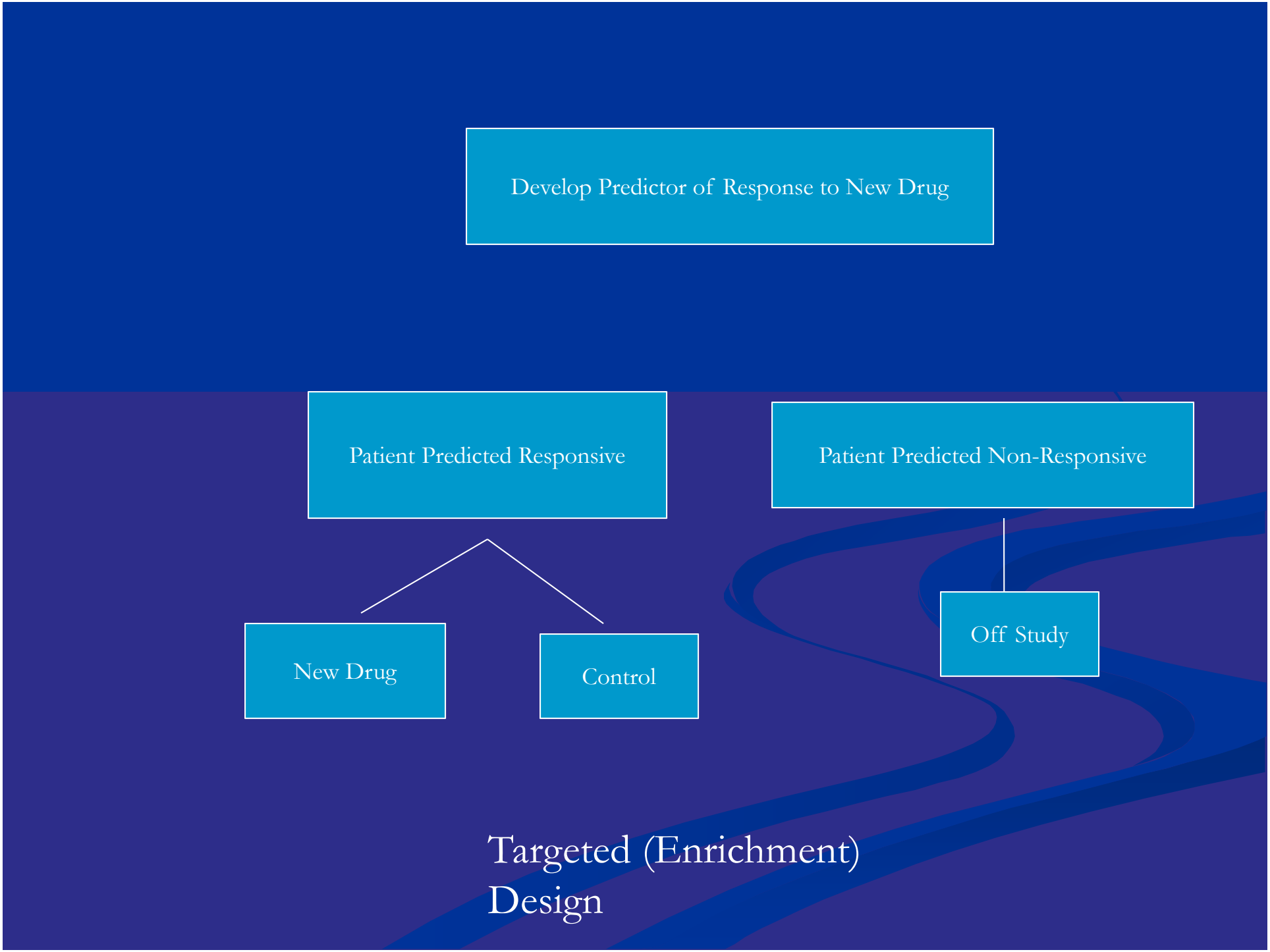
Modern Tumor Biology

- Tumors of a primary site differ with regard to the mutations which cause them, natural history and response to therapy
- Molecularly targeted drugs are likely to be effective only for tumors that are driven by de-regulation of the pathway which is a target of the drug

- The traditional broad eligibility clinical trial is inappropriate in these cases. It leads to treating patients with drugs to which we don't expect them to benefit and to doing analyses that are in conflict with good science.

When the Biology is Clear the Development Path is Straightforward

- Develop a classifier that identifies the patients likely to benefit from the new drug
- Develop an analytically validated test
  - Measures what it should accurately and reproducibly
- Design a focused clinical trial to evaluate effectiveness of the new treatment in test + patients



Targeted (Enrichment) Design

Evaluating the Efficiency of Targeted Design

- Simon R and Maitnourim A. Evaluating the efficiency of targeted designs for randomized clinical trials. Clinical Cancer Research 10:6759-63, 2004; Correction and supplement 12:3229, 2006
- Maitnourim A and Simon R. On the efficiency of targeted clinical trials. Statistics in Medicine 24:329-339, 2005.

RandRat = $n_{untargeted}/n_{targeted}$

$$RandRat \approx \left(\frac{T_+}{p_+ T_+ + (1 - p_+) T_-} \right)^2$$

- If $T_- = 0$, RandRat = $1/\ p_+^2$
  - if $p_+ = 0.5$, RandRat=4
- If $T_- = T_+/2$, RandRat = $4/(p_+ + 1)^2$
  - if $p_+ = 0.5$, RandRat=16/9=1.77

Comparing T vs C on Survival or DFS

5% 2-sided Significance and 90% Power

% Reduction in Hazard	Number of Events Required
25%	509
30%	332
35%	227
40%	162
45%	118
50%	88

Successful use of targeted enrichment design

- Trastuzumab, pertuzumab, ado-trastuzumab emtansine for HER2 over-expressed or amplified breast cancer
- Vemurafinib, dabrafinib, trametinib for BRAF mutated melanoma
- Crizotinib and ceritinib in ALK translocated NSCLC
- Afatinib in EGFR mutated NSCLC

Regulatory Pathway for Test

- Companion diagnostic test with intended use of identifying patients who have disease subtype for which the drug is proven effective

Advantages of enrichment design

- Targets larger treatment effect less diluted by non-sensitive tumors
- Avoids exposing patients less likely to benefit to adverse effects of drug until drug is shown effective for those whom it is supposed to benefit
- Clarity of interpretation

- If the drug is effective in test positive patients, it can be later evaluated in test negative patients.
  - Saves test – patients toxicity until drug is shown effective in the target population it should work in

All comers design

- Invites poor design
  - Too few test + patients
  - Too many test – patients
  - Failure to have a specific analysis plan
- Invites inappropriate analysis
  - Inappropriate requirement of not doing subset analysis unless ITT test is significant and interaction is significant

$$\mathbf{RandRat = n_{untargeted}/n_{targeted}}$$

$$RandRat \approx \left(\frac{T_+}{p_+ B_+ + (1 - p_+) B_-} \right)^2$$

- B_+=TE in biology + pts
- T_+=TE in test + pts
- T_+=ppvB_+ + (1-ppv)B_-

$$ppv = \left\{ 1 + \frac{1 - spec}{sens} \frac{1 - p_+}{p_+} \right\}^{-1}$$

Sensitivity	Specificity	p+	B-	PPV	Rand Ratio
0.8	0.8	0.33	0	0.67	2
0.8	0.8	0.33	B+/2	0.67	1.25

Run-In Phase III Trial Design With Pharmacodynamics Predictive Biomarkers

Fangxin Hong, Richard Simon

Manuscript received January 2, 2013; revised July 31, 2013; accepted August 1, 2013.

Correspondence to: Richard Simon, PhD, Biometric Research Branch, National Cancer Institute, Bethesda, MD, 20892 (e-mail: rsimon@mail.nih.gov).

Background Developments in biotechnology have stimulated the use of predictive biomarkers to identify patients who are likely to benefit from a targeted therapy. Several randomized phase III designs have been introduced for development of a targeted therapy using a diagnostic test. Most such designs require biomarkers measured before treatment. In many cases, it has been very difficult to identify such biomarkers. Promising candidate biomarkers can sometimes be effectively measured after a short run-in period on the new treatment.

Methods We introduce a new design for phase III trials with a candidate predictive pharmacodynamic biomarker measured after a short run-in period. Depending on the therapy and the biomarker performance, the trial would either randomize all patients but perform a separate analysis on the biomarker-positive patients or only randomize marker-positive patients after the run-in period. We evaluate the proposed design compared with the conventional phase III design and discuss how to design a run-in trial based on phase II studies.

Results The proposed design achieves a major sample size reduction compared with the conventional randomized phase III design in many cases when the biomarker has good sensitivity (≥0.7) and specificity (≥0.7). This requires that the biomarker be measured accurately and be indicative of drug activity. However, the proposed design loses some of its advantage when the proportion of potential responders is large (>50%) or the effect on survival from run-in period is substantial.

Conclusions Incorporating a pharmacodynamic biomarker requires careful consideration but can expand the capacity of clinical trials to personalize treatment decisions and enhance therapeutics development.

J Natl Cancer Inst

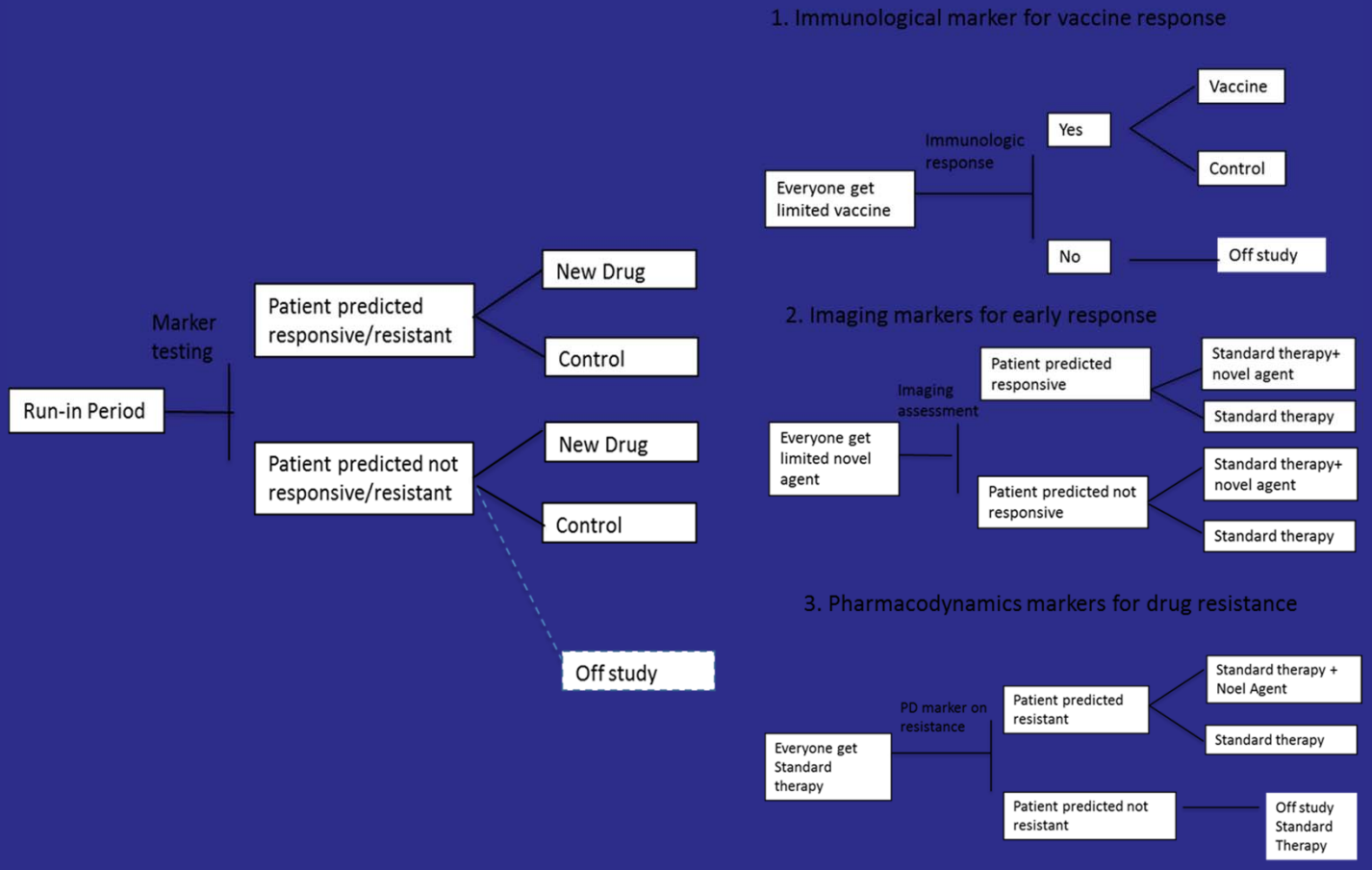
Improved understanding of cancer biology has stimulated the development of molecularly targeted cancer treatments that will likely only benefit patients whose tumors are driven by deregulation of the drug targets. The standard phase III trial testing average drug effect across patients with broad eligibility criteria is often no longer efficient. Even when such trials result in statistical significance, a large proportion of the patients do not benefit from the new treatment.

A key component in developing targeted therapy is the identification of predictive biomarkers that can identify patients who are likely to benefit. Effective predictive biomarkers can benefit patients, control costs by personalizing treatment, and enhance the efficiency of clinical development. Statisticians are challenged to develop new designs and analysis strategies to incorporate predictive biomarkers. Several randomized phase III designs have been previously introduced for this purpose (1,2), including the marker strategy design, the enrichment design (3), and the marker-stratified design (4). All of these designs require pretreatment biomarker measurement. In many cases, it has been very difficult to identify such pretreatment biomarkers. Biomarkers measured after receiving the randomized treatment are generally not suitable because different treatment arms could have differential effects on biomarker values. Some studies use posttreatment biomarkers as surrogates of clinical outcome, but establishing an intermediate endpoint as a valid surrogate is quite difficult (5).

Run-in periods in which all patients receive the test drug for a short period of time have been used in some clinical trials to exclude or select patients for subsequent randomization (6). The earliest run-in designs were implemented to exclude patients with poor compliance to treatment (7,8). Run-in periods in which all patients receive placebo have been used to exclude placebo responders (9).

In this article, we explore the use of pharmacodymaic biomarkers measured after a short run-in period on the new treatment as a predictive biomarker. A wide variety of such biomarkers are potentially available. Immunologic response to a therapeutic cancer vaccine is one example. Dendritic cell–based cancer vaccines, although expensive, are very effective for inducing antitumor immunity in a variety of cancers (10,11). However, clinical responses are observed in only a subset of patients (12). Assessing early immunologic response may efficiently identify the subset of patients who will have a greater chance of eventually having clinical responses. A second area is the use of mechanistic markers. Downregulation

Run-in Designs
Fangxin Hong & R Simon



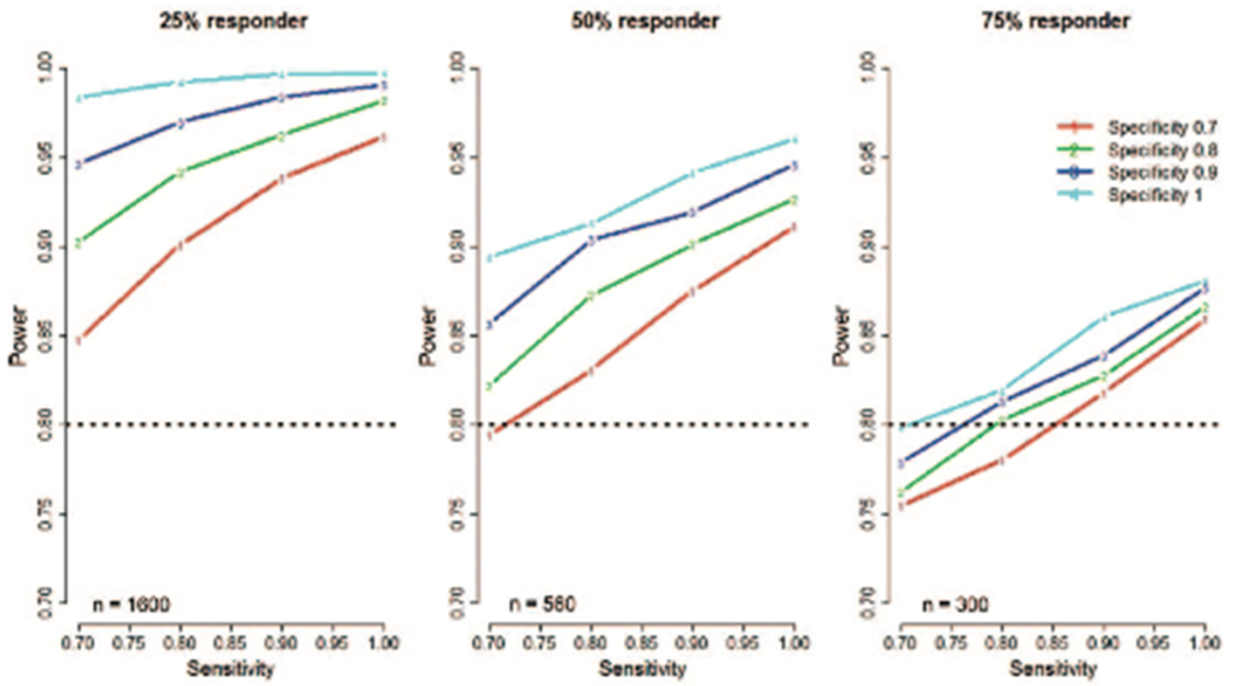


Figure 2. With sample sizes that give 80% power for the standard design, the trial-level power with the run-in design (**solid lines**) is shown when randomizing all patients, for a series of sensitivity and specificity of the biomarker, under 25%, 50%, and 75% prevalence** of true responders, with no run-in effect ($\theta_0 = 1$).

When the biology is not so clear

- It is difficult to have the right single completely defined predictive biomarker identified and analytically validated by the time the pivotal trial of a new drug is ready to start accrual

Biostatistics Advance Access published March 21, 2013

Biostatistics (2013), *pp.* 1–13
doi:10.1093/biostatistics/kxt010

Adaptive enrichment designs for clinical trials

NOAH SIMON*

Department of Statistics, Stanford University, Stanford, CA 94305, USA
nsimon@stanford.edu

RICHARD SIMON

Biometric Research Branch, National Cancer Institute, Bethesda, MD 20892, USA

SUMMARY

Modern medicine has graduated from broad spectrum treatments to targeted therapeutics. New drugs recognize the recently discovered heterogeneity of many diseases previously considered to be fairly homogeneous. These treatments attack specific genetic pathways which are only dysregulated in some smaller subset of patients with the disease. Often this subset is only rudimentarily understood until well into large-scale clinical trials. As such, standard practice has been to enroll a broad range of patients and run post hoc subset analysis to determine those who may particularly benefit. This unnecessarily exposes many patients to hazardous side effects, and may vastly decrease the efficiency of the trial (especially if only a small subset of patients benefit). In this manuscript, we propose a class of adaptive enrichment designs that allow the eligibility criteria of a trial to be adaptively updated during the trial, restricting entry to patients likely to benefit from the new treatment. We show that our designs both preserve the type 1 error, and in a variety of cases provide a substantial increase in power.

Keywords: Adaptive clinical trials; Biomarker; Cutpoint; Enrichment.

1. INTRODUCTION

The literature on adaptive clinical trial design has focused on sample-size reestimation, changing the plan for interim analyses, or modifying randomization weights (Chow and Chang, 2007; Muller and Schafer, 2001; Rosenberger and Lachin, 1993; Karrison *and others*, 2003; Kim *and others*, 2011). In oncology therapeutics development, attention has turned toward discovery of baseline predictive biomarkers to identify patients likely to benefit from the new treatment (Papadopoulos *and others*, 2006; Schilsky, 2007; Sawyers, 2008). Tumors of most body sites have been found to be biologically heterogeneous with regard to their causal mutations and molecularly targeted drugs are unlikely to benefit most patients in the broad diagnostic categories traditionally included in clinical trials. When the pathophysiology of the disease and the mechanism of action of the drug are well understood, a binary predictive biomarker can be identified prior to or early in clinical development and used to restrict entry of patients to the pivotal phase 3 clinical trials comparing the new drug with a suitable control. Such "enrichment" designs can serve to magnify

*To whom correspondence should be addressed.

- Provides a general framework for adaptive enrichment, i.e. restricting the eligibility criteria during the course of the trial based on interim results.
- Framework includes threshold based enrichment or enrichment based on multi-marker modeling
- Framework handles multiple types of endpoints (continuous, binary, time-to-event)

- One primary statistical significance test is performed at the end of the trial, including all randomized patients, of the strong null hypothesis that the new treatment is uniformly ineffective
- Framework identifies classes of significance tests which preserve the type I error even with time dependent and data dependent changes to outcome distributions of patients

Simulation of adaptive threshold enrichment

- Single biomarker uniformly distributed on (0,1)
- K candidate thresholds
- Binary response with probability $p_0(b)$ or $p_1(b)$
- True threshold x^*
- $p_0(b)=p_0$ for all b
- $p_1(b)=p_0$ for $b<x^*$, p_1 for $b>x^*$
- Single interim analysis

At interim analysis compute mle of $\{p_0, p_1, x^*\}$ subject to $p_0 \leq p_1$

Perform futility analysis: If this maximized log likelihood does not exceed null log likelihood with no treatment effect by at least 0.25, terminate trial.

Otherwise accrual was restricted to those with $b \geq \hat{x}^*$ for remainder of trial.

N total patients, n_1 at interim analysis.

Test statistic $T=\sum \{y_i z_i + (1-y_i)(1-z_i)\}$

$z_i = (0,1)$ treatment indicator

$y_i = (0,1)$ response

p_0=.2, p_1=.5, K=5, N_{tot}=200, all pts 100/yr

True cut-point	Power adaptive	Power non-adaptive	Accrual adaptive	Accrual non-adaptive
.25	.968	.955	2.55	2.25
.5	.897	.726	3.19	3.25
.67	.768	.424	3.97	4.75

Test statistic $T=\sum \{y_i z_i + (1-y_i)(1-z_i)\}$

$z_i = (0,1)$ treatment indicator

$y_i = (0,1)$ response

$E[T|\underline{y}]=\sum (0.5y_i + 0.5(1-y_i)) = n/2$

Significance tests that preserve type I error with group sequential adaption

t_k = treatment effect statistic based on all patients accrued in block k

L_k = all data (outcomes, covariates, treatment assignments) for blocks 1,...,k

We require that the $\{t_k\}$ be defined so that under null hypothesis,
the distribution of t_k is known and independent of L_{k-1} for $k = 1,...,K$

Then any test statistic $G(t_1,...,t_K)$ which is a function of the data only thru $\{t_k\}$
has a known null distribution and the type I error is preserved.

e.g. $G = \sum_{k=1}^{K} w_k t_k$

Single binary marker with two stage design

- Total sample size N patients
- At interim analysis decide
  - Whether to terminate accrual of M– patients and continue accrual of M+ till total sample size of N. Target population will be M+ patients
  - Whether to continue accrual of marker – patients and target population will be union of M+ and M-
  - Whether to terminate accrual of M+ and M- and accept null hypothesis

Biometrika (2011), **98**, 4, *pp.* 845–860 doi: 10.1093/biomet/asr055

Printed in Great Britain

Optimizing randomized trial designs to distinguish which subpopulations benefit from treatment

BY M. ROSENBLUM

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe St., Room E3616, Baltimore, Maryland 21205, U.S.A.

mrosenbl@jhsph.edu

AND M. J. VAN DER LAAN

Division of Biostatistics, University of California, School of Public Health, 101 Haviland Hall, Berkeley, California 94720-7358, U.S.A.

laan@stat.berkeley.edu

SUMMARY

It is a challenge to evaluate experimental treatments where it is suspected that the treatment effect may only be strong for certain subpopulations, such as those having a high initial severity of disease, or those having a particular gene variant. Standard randomized controlled trials can have low power in such situations. They also are not optimized to distinguish which subpopulations benefit from a treatment. With the goal of overcoming these limitations, we consider randomized trial designs in which the criteria for patient enrollment may be changed, in a preplanned manner, based on interim analyses. Since such designs allow data-dependent changes to the population enrolled, care must be taken to ensure strong control of the familywise Type I error rate. Our main contribution is a general method for constructing randomized trial designs that allow changes to the population enrolled based on interim data using a prespecified decision rule, for which the asymptotic, familywise Type I error rate is strongly controlled at a specified level α. As a demonstration of our method, we prove new, sharp results for a simple, two-stage enrichment design. We then compare this design to fixed designs, focusing on each design's ability to determine the overall and subpopulation-specific treatment effects.

Some key words: Adaptive design; Enrichment design; Group sequential design; Optimization; Patient-oriented research; Randomized trial; Subpopulation.

Adaptive patient enrichment designs in therapeutic trials

Sue-Jane Wang[*,1], **H. M. James Hung**[2], and **Robert T. O'Neill**[1]

[1] Office of Biostatistics, Division of Biometrics I/OB, Office of Translational Sciences, Center for Drug Evaluation and Research, US Food and Drug Administration, USA
[2] Division of Biometrics I/OB, Office of Translational Sciences, Center for Drug Evaluation and Research, US Food and Drug Administration, USA

Received 15 April 2008, revised 3 January 2009, accepted 9 January 2009

The utility of clinical trial designs with adaptive patient enrichment is investigated in an adequate and well-controlled trial setting. The overall treatment effect is the weighted average of the treatment effects in the mutually exclusive subsets of the originally intended entire study population. The adaptive enrichment approaches permit assessment of treatment effect that may be applicable to specific nested patient (sub)sets due to heterogeneous patient characteristics and/or differential response to treatment, *e.g.* a responsive patient subset *versus* a lack of beneficial patient subset, in all patient (sub)sets studied. The adaptive enrichment approaches considered include three adaptive design scenarios: (i) total sample size fixed and with futility stopping, (ii) sample size adaptation and futility stopping, and (iii) sample size adaptation without futility stopping. We show that regardless of whether the treatment effect eventually assessed is applicable to the originally studied patient population or only to the nested patient subsets; it is possible to devise an adaptive enrichment approach that statistically outperforms one-size-fits-all fixed design approach and the fixed design with a pre-specified multiple test procedure. We emphasize the need of additional studies to replicate the finding of a treatment effect in an enriched patient subset. The replication studies are likely to need fewer number of patients because of an identified treatment effect size that is larger than the diluted overall effect size. The adaptive designs, when applicable, are along the line of efficiency consideration in a drug development program.

Key words: Adaptive enrichment algorithm; Futility; Nested patient subset; Strong control of experiment-wise type I error; Weighted Z-statistic.

Supporting Information for this article is available from the author or on the WWW under http://dx.doi.org/10.1002/bimj.200900003

Research Article

Statistics
in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

A Two-stage Bayesian Design for Co-Development of New Drugs and Companion Diagnostics

Stella Wanjugu Karuri [a] and Richard Simon[b]*

Most new drug development in oncology is based on targeting specific molecules. Genomic profiles and deregulated drug targets vary from patient to patient making new treatments likely to benefit only a subset of patients traditionally grouped in the same clinical trials. Predictive biomarkers are being developed to identify patients who are most likely to benefit from a particular treatment; however their biological basis is not always conclusive. The inclusion of marker negative patients in a trial is therefore sometimes necessary for a more informative evaluation of the therapy. In this paper we present a two-stage Bayesian design which includes both marker positive and marker negative patients in a clinical trial. We formulate a family of prior distributions that represent the degree of a-priori confidence in the predictive biomarker. To avoid exposing patients to a treatment to which they may not be expected to benefit, an interim analysis is performed which may stop accrual of marker negative patients or accrual of all patients. We demonstrate with simulations that the design and priors used control type I errors, give adequate power and enable the early futility analysis of test negative patients to be based on prior specification on the strength of evidence in the biomarker.

Keywords: Clinical trials design, predictive biomarkers, bayesian inference, prior distribution, Type I error probabilities

δ_+ = treatment effect in test + patients

δ_- = treatment effect in test - patients

Two-point priors for δ_+ and δ_- with values $\{0, \delta^*\}$

Approximate survival analysis with

$$(\hat{\delta}_+, \hat{\delta}_-) : \; N\left((\delta_+, \delta_-), \begin{pmatrix} 4/E_+ & 0 \\ 0 & 4/E_- \end{pmatrix} \right)$$

$$\Pr[\delta_+ = \delta_- = 0] = p_{00}$$

$$\Pr[\delta_- = 0 \mid \delta_+ = \delta^*] = r_1$$

$$\Pr[\delta_+ = 0 \mid \delta_- = \delta^*] = r_2$$

Strong confidence in test: large r_1
Weak confidence in test: small r_1
p_{00} selected to control type I error rates

$$\Pr[\delta_+ = \delta_- = 0] = p_{00} = 0.1$$

$$\Pr[\delta_- = 0 \mid \delta_+ = \delta^*] = r_1 = 0.1 - 0.9$$

$$\Pr[\delta_+ = 0 \mid \delta_- = \delta^*] = r_2 = 0.1$$

Strong confidence in test: Small r_2 and large r_1
Weak confidence in test: Small r_2 and small r_1
p_{00} selected to control type I error rates

Interim Analysis

Terminate accrual of test - patients if $\Pr[\delta_- = 0 \mid \hat{\delta}_+^{(1)}, \hat{\delta}_-^{(1)}] > \gamma$

Terminate accrual to trial if $\Pr[\delta_+ = 0 \mid \hat{\delta}_+^{(1)}, \hat{\delta}_-^{(1)}] > \gamma$

Final Analysis
Probabilistic Indication
Classifier

Compute joint 4-point posterior distribution using full clinical trial dataset

$\Pr[\delta_+, \delta_- \mid \hat{\delta}_+^{(2)}, \hat{\delta}_-^{(2)}]$ and calculate marginals

Probability new treatment benefits test - patients $\Pr[\delta_- = \delta^* \mid \hat{\delta}_+^{(2)}, \hat{\delta}_-^{(2)}]$

Probability new treatment benefits test + patients $\Pr[\delta_+ = \delta^* \mid \hat{\delta}_+^{(2)}, \hat{\delta}_-^{(2)}]$

$$P(\delta_+ = 0|\hat{\delta}_+, \hat{\delta}_-) = \left(1 + \frac{aq_1 L_{\delta_*,0} + aL_{\delta_*,\delta_*}}{q_{00} L_{0,0} + aq_2 L_{0,\delta_*}}\right)^{-1}, \tag{7}$$

$$P(\delta_- = 0|\hat{\delta}_+, \hat{\delta}_-) = \left(1 + \frac{aq_2 L_{0,\delta_*} + aL_{\delta_*,\delta_*}}{q_{00} L_{0,0} + aq_1 L_{\delta_*,0}}\right)^{-1}, \tag{8}$$

$$P(\delta_+ = 0, \delta_- = 0|\hat{\delta}_+, \hat{\delta}_-) = \left(1 + \frac{aq_2 L_{0,\delta_*} + aq_1 L_{\delta_*,0} + aL_{\delta_*,\delta_*}}{q_{00} L_{0,0}}\right)^{-1}, \tag{9}$$

where $a = (1 - r_1)(1 - r_2)/(1 - r_1 r_2)$, $q_1 = r_1/(1 - r_1)$, $q_2 = r_2/(1 - r_2)$, $q_{00} = p_{00}/(1 - p_{00})$, and $L_{i,j} = P(\hat{\delta}_+, \hat{\delta}_-|\delta_+ = i, \delta_- = j)$ denote the marginal conditional probabilities of the bivariate Normal density given in Equation (3). Complete *a priori* confidence in the utility of the classifier means

"Adaptive Final Analysis Plans"

- Adaptive signature design
- Cross-validated adaptive signature design

Adaptive Signature Design: An Adaptive Clinical Trial Design for Generating and Prospectively Testing A Gene Expression Signature for Sensitive Patients

Boris Freidlin and Richard Simon

Abstract **Purpose:** A new generation of molecularly targeted agents is entering the definitive stage of clinical evaluation. Many of these drugs benefit only a subset of treated patients and may be overlooked by the traditional, broad-eligibility approach to randomized clinical trials. Thus, there is a need for development of novel statistical methodology for rapid evaluation of these agents.

Experimental Design: We propose a new adaptive design for randomized clinical trials of targeted agents in settings where an assay or signature that identifies sensitive patients is not available at the outset of the study. The design combines prospective development of a gene expression – based classifier to select sensitive patients with a properly powered test for overall effect.

Results: Performance of the adaptive design, relative to the more traditional design, is evaluated in a simulation study. It is shown that when the proportion of patients sensitive to the new drug is low, the adaptive design substantially reduces the chance of false rejection of effective new treatments. When the new treatment is broadly effective, the adaptive design has power to detect the overall effect similar to the traditional design. Formulas are provided to determine the situations in which the new design is advantageous.

Conclusion: Development of a gene expression – based classifier to identify the subset of sensitive patients can be prospectively incorporated into a randomized phase III design without compromising the ability to detect an overall effect.

Developments in tumor biology have resulted in shift toward molecularly targeted drugs (1–3). Most human tumor types are heterogeneous with regard to molecular pathogenesis, genomic signatures, and phenotypic properties. As a result, only a subset of the patients with a given cancer is likely to benefit from a targeted agent (4). This complicates all stages of clinical development, especially randomized phase III trials (5, 6). In some cases, predictive assays that can accurately identify patients who are likely to benefit from the new therapy have been developed. Then, targeted randomized designs that restrict eligibility to patients with sensitive tumors should be used (7). However, reliable assays to select sensitive patients are often not available (8, 9). Consequently, traditional randomized clinical trails with broad eligibility criteria are routinely used to evaluate such agents. This is generally inefficient and may lead to missing effective agents.

Genomic technologies, such as microarrays and single nucleotide polymorphism genotyping, are powerful tools that hold a great potential for identifying patients who are likely to benefit from a targeted agent (10, 11). However, due to the large number of genes available for analysis, interpretation of these data is complicated. Separation of reliable evidence from the random patterns inherent in high-dimensional data requires specialized statistical methodology that is prospectively incorporated in the trial design. Practical implementation of such designs has been lagging. In particular, analysis of microarray data from phase III randomized studies is usually considered secondary to the primary overall comparison of all eligible patients. Many analyses are not explicitly written into protocols and done retrospectively, mainly as "hypothesis-generating" tools.

We propose a new adaptive design for randomized clinical trials of molecularly targeted agents in settings where an assay or signature that identifies sensitive patients is not available. Our approach includes three components: (*a*) a statistically valid identification, based on the first stage of the trial, of the subset of patients who are most likely to benefit from the new agent; (*b*) a properly powered test of overall treatment effect at the end of the trial using all randomized patients; and (*c*) a test of treatment effect for the subset identified in the first stage, but using only patients randomized in the remainder of the trial. The components are prospectively incorporated into a single phase III randomized clinical trial with the overall false-positive error rate controlled at a prespecified level.

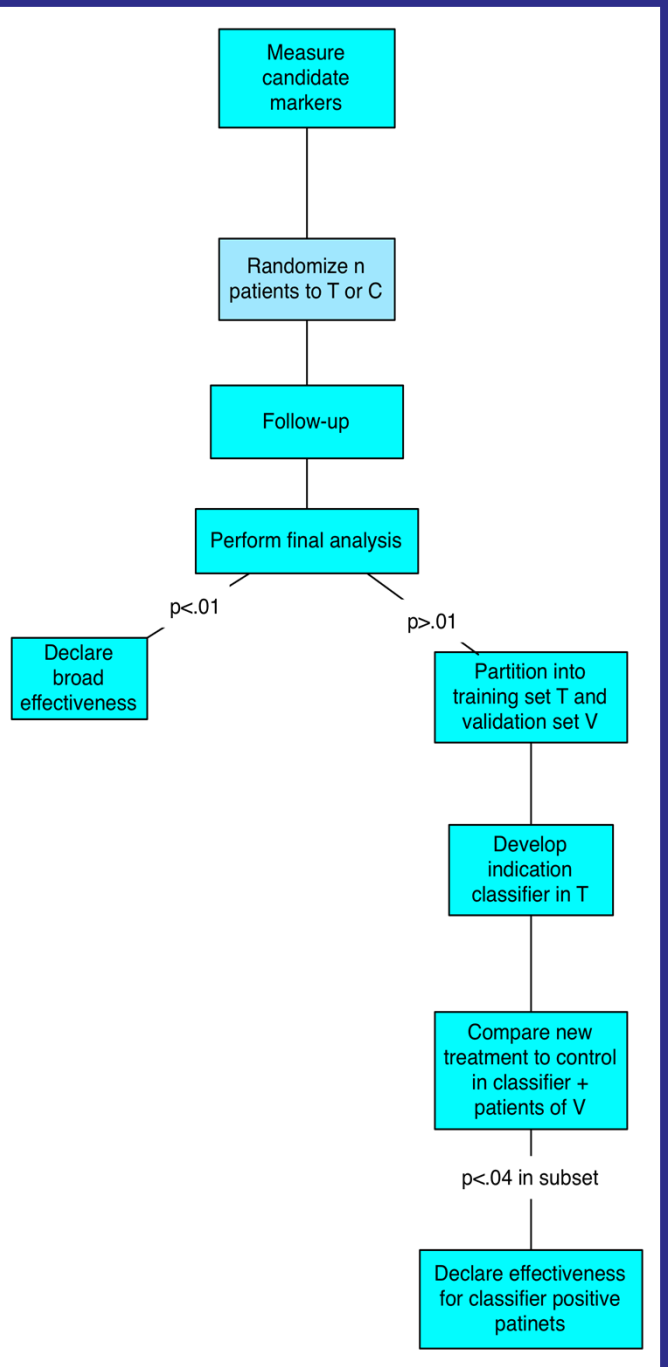
Authors' Affiliation: Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, Maryland

Received 3/18/05; revised 7/18/05; accepted 8/4/05.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Requests for reprints: Boris Freidlin, Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, 6130 Executive Boulevard, EPN 8122, MSC 7434, Bethesda, MD 20892-7434. Phone: 301-402-0640; Fax: 301-402-0560; E-mail: freidlinb@ctep.nci.nih.gov.

doi:10.1158/1078-0432.CCR-05-0605


Measure candidate markers
Randomize n patients to T or C
Follow-up
Perform final analysis
p<.01
p>.01
Declare broad effectiveness
Partition into training set T and validation set V
Develop indication classifier in T
Compare new treatment to control in classifier + patients of V
p<.04 in subset
Declare effectiveness for classifier positive patinets

- The indication classifier is not a binary classifier of whether a patient has good prognosis or poor prognosis
- It is a “two sample classifier” of whether the prognosis of a patient on E is better than the prognosis of the patient on C

- The indication classifier maps the vector of candidate covariates into {E,C} indicating which treatment is predicted superior for that patient
- The classifier need not use all the covariates but variable selection must be determined using only the training set
  - Variable selection may be based on selecting variables with apparent interactions with treatment, with cut-off for variable selection determined by cross-validation within training set for optimal classification
- The indication classifier can be a probabilistic classifier

For high dimension

$$\log\left(\lambda(t,\underline{x},z)/\lambda_0(t)\right)=\alpha z+(1-z)\underline{\beta}'\underline{x}+z\underline{\eta}'\underline{x}$$

z=(0,1) treatment indicator

Use penalized proportional hazard model

to obtain estimates $\hat{\underline{\beta}},\hat{\underline{\eta}}$

Treatment effect function $\alpha+\eta' x-\beta' x$

Classify case as likely to benefit from E if

$$\Delta(\underline{x})=\hat{\alpha}+\hat{\underline{\eta}}'\underline{x}-\hat{\underline{\beta}}'\underline{x}\le\Delta^*$$

1. $\Delta^* = \log(0.6)$

2. $\Delta^* =$ 33rd percentile of $\alpha + (\hat{\beta} - \hat{\eta})' x$ in training set

3. Optimize by cross-validation in training set

For low dimension

$$\hat{\alpha} + \hat{\underline{\gamma}}' \underline{x} : \; N(\alpha + \underline{\gamma}' \underline{x}, \sigma^2(\underline{x}))$$

$$\sigma^2(x) = (1, \underline{x})' \Sigma (1, \underline{x})$$

$$\Pr\left[\text{E preferred}\right] = \Pr\left[\alpha + \underline{\gamma}' \underline{x} \le \Delta\right]$$

$$= \Pr\left[\frac{\left(\hat{\alpha} + \hat{\underline{\gamma}}' \underline{x}\right) - \left(\alpha + \underline{\gamma}' \underline{x}\right)}{\sigma(\underline{x})} \ge \frac{\left(\hat{\alpha} + \hat{\underline{\gamma}}' \underline{x}\right) - \Delta}{\sigma(\underline{x})}\right]$$

$$; \; \Phi\left\{\frac{\Delta - \left(\hat{\alpha} + \hat{\underline{\gamma}}' \underline{x}\right)}{\hat{\sigma}(\underline{x})}\right\}$$

Recommend E if

$\Pr[\text{E preferred}] \geq 0.8$

True Model

$$\log\left(\frac{p_i}{1-p_i}\right) = \mu^* + \lambda^* t_i + t_i \underline{\gamma}^* (x_{i1}, \ldots, x_{iK})$$

Classifier Development

- Using data from stage 1 patients, fit all single gene logistic models (j=1,…,M)
- Select genes with interaction significant at level α

$$\log it(p_i) = \mu + \lambda_j t_i + \nu_j x_{ij} + \beta_j t_i x_{ij}$$

Classification of Stage 2 Patients

- For i'th stage 2 patient, selected gene j votes to classify patient as preferentially sensitive to T if

$$\exp\left\{\hat{\lambda}_j + \hat{\beta}_j x_{ij}\right\} > R$$

Classification of Stage 2 Patients

- Classify i'th stage 2 patient as differentially sensitive to E relative to C if at least G selected genes vote for differential sensitivity of that patient

Treatment effect restricted to subset.
10% of patients sensitive, 10 sensitivity genes, 10,000 genes, 400 patients.

Test	Power
Overall .05 level test	46.7
Overall .04 level test	43.1
Sensitive subset .01 level test (performed only when overall .04 level test is negative)	42.2
Overall adaptive signature design	85.3

Overall treatment effect, no subset effect.
10% of patients sensitive, 10 sensitivity genes, 10,000 genes, 400 patients.

Test	Power
Overall .05 level test	74.2
Overall .04 level test	70.9
Sensitive subset .01 level test	1.0
Overall adaptive signature design	70.9

Key Idea

- Replace multiple significance testing by development of one indication classifier and obtain unbiased estimates of the properties of that classifier if used on future patients

Special Report

Clinical Cancer Research

Adaptive Clinical Trial Designs for Simultaneous Testing of Matched Diagnostics and Therapeutics

Howard I. Scher[1], Shelley Fuld Nasso[2], Eric H. Rubin[3], and Richard Simon[4]

Abstract

A critical challenge in the development of new molecularly targeted anticancer drugs is the identification of predictive biomarkers and the concurrent development of diagnostics for these biomarkers. Developing matched diagnostics and therapeutics will require new clinical trial designs and methods of data analysis. The use of adaptive design in phase III trials may offer new opportunities for matched diagnosis and treatment because the size of the trial can allow for subpopulation analysis. We present an adaptive phase III trial design that can identify a suitable target population during the early course of the trial, enabling the efficacy of an experimental therapeutic to be evaluated within the target population as a later part of the same trial. The use of such an adaptive approach to clinical trial design has the potential to greatly improve the field of oncology and facilitate the development of personalized medicine. *Clin Cancer Res; 17(21); 6634–40. ©2011 AACR.*

Introductory Note

At the 2010 Conference on Clinical Cancer Research, co-convened by Friends of Cancer Research and the Engelberg Center for Health Care Reform at the Brookings Institution, participants explored 4 pressing challenges in the field. Articles summarizing the panel's recommendations on each of these topics are featured in this issue of *Clinical Cancer Research* (1–4).

Key Role of Companion Diagnostics in Oncology Drug Development

Nearly all cancer drugs being developed today are designed to inhibit molecular targets that have been identified as being dysregulated in human tumors. Genomics has established that the dysregulated pathways and mutated genes in tumors originating in a particular primary site are highly variable. To optimally evaluate and utilize a targeted approach requires the concurrent development of diagnostics that enable the identification of those tumors that are most likely to be sensitive to the anticancer effects of a particular drug or drug combination. The reality of codeveloping a matched diagnostic and therapeutic has profound implications for the clinical trial designs used in drug development. Trials of cytotoxic drugs typically enroll unselected patients at a particular point in the continuum of a disease in the hope that the response of tumors that are sensitive to the treatment will be sufficient to show benefit for the population as a whole. Although this approach may lead to broad labeling indications, it also exposes patients with nonsensitive tumors to unnecessary toxicities and increases the possibility of discarding a drug that may dramatically benefit a subset of patients. Consequently, this strategy is not viable for molecularly targeted agents, in which the activity is likely to be restricted and determined more by the genomic alteration(s) within a tumor at the time treatment is being considered than by the primary site in which the tumor originated. The use of anatomically based (i.e., primary site of disease), "all comers" approaches to develop targeted approaches has typically led to failure in phase III studies, or demonstration of "success" based on statistically significant but clinically questionable benefits (5).

Although developing the right drug for a specific patient has great value to the individual and is critical for controlling the costs of health care, it dramatically increases the complexity of the drug development process. For many drugs, the complexities of identifying a predictive biomarker and the practical complexities of developing analytically valid diagnostic tests for the biomarker are grossly underestimated. Knowing when to start the development of the diagnostic is also an issue, particularly when the effectiveness of the drug in any population is uncertain. Developing the right drug for the right subset of patients requires new clinical trial designs and new paradigms of data analysis.

Efforts to codevelop a matched diagnostic and therapeutic face other challenges as well. Even with extensive

Authors' Affiliations: [1]Department of Medicine, Genitourinary Oncology Service, Memorial Sloan-Kettering Cancer Center, and Weill Cornell Medical College, New York, New York; [2]Susan G. Komen for the Cure Advocacy Alliance, Washington, District of Columbia; [3]Oncology Clinical Research, Merck Research Laboratories, Whitehouse Station, New Jersey; and [4]Biometric Research Branch, National Cancer Institute, Bethesda, Maryland

Corresponding Author: Howard I. Scher, Department of Medicine, Sidney Kimmel Center for Prostate and Urologic Cancers, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, NY. Phone: 646-422-4323; Fax: 212-988-0851; E-mail: scherh@mskcc.org

doi: 10.1158/1078-0432.CCR-11-1105

Cancer Therapy: Clinical

Clinical Cancer Research

The Cross-Validated Adaptive Signature Design

Boris Freidlin[1], Wenyu Jiang[2], and Richard Simon[1]

Abstract

Purpose: Many anticancer therapies benefit only a subset of treated patients and may be overlooked by the traditional broad eligibility approach to design phase III clinical trials. New biotechnologies such as microarrays can be used to identify the patients that are most likely to benefit from anticancer therapies. However, due to the high-dimensional nature of the genomic data, developing a reliable classifier by the time the definitive phase III trail is designed may not be feasible.

Experimental Design: Previously, Freidlin and Simon (*Clinical Cancer Research*, 2005) introduced the adaptive signature design that combines a prospective development of a sensitive patient classifier and a properly powered test for overall effect in a single pivotal trial. In this article, we propose a cross-validation extension of the adaptive signature design that optimizes the efficiency of both the classifier development and the validation components of the design.

Results: The new design is evaluated through simulations and is applied to data from a randomized breast cancer trial.

Conclusion: The cross-validation approach is shown to considerably improve the performance of the adaptive signature design. We also describe approaches to the estimation of the treatment effect for the identified sensitive subpopulation. *Clin Cancer Res; 16(2); 691–8. ©2010 AACR.*

Due to the molecular heterogeneity of most human cancers, only a subset of treated patients benefit from a given therapy. This is particularly relevant for the new generation of anticancer agents that target specific molecular pathways (1–3). Genomic (or proteinomic) technologies such as microarrays provide powerful tools for identifying a genetic signature (diagnostic test) for patients who are most likely to benefit from a targeted agent. Ideally, such diagnostic test should be developed and validated before commencing the definitive phase III trial (4). However, due to the complexity of signaling pathways and the large number of genes available for analysis, the development of a reliable diagnostic classifier using early nonrandomized phase II data is often not feasible. Conducting a phase III randomized clinical trial (RCT) requires considerable time and resources. Therefore, clinical trial designs that allow combining the definitive evaluation of a new agent with the development of the companion diagnostic test can considerably speed up the introduction of new cancer therapies.

Previously, the adaptive signature design (ASD) has been proposed for settings where a signature to identify sensitive patients is not available (5). The design combines the prospective development of a pharmacogenomic diagnostic test (signature) to select sensitive patients with a properly powered test for overall effect. It was shown that when the proportion of patients sensitive to the new drug is low, the ASD substantially reduces the chance of false rejection of effective new treatments. When the new treatment is broadly effective, the power of the adaptive design to detect the overall effect is similar to that of the traditional design.

The signature component of the ASD carries out signature development and validation on the mutually exclusive subgroups of patients (e.g., half of the study population is used to develop a signature and another half to validate it). Although the conceptual simplicity of this approach is appealing, it also limits its power as only half of the patients are used for signature development and half for validation. This is especially relevant in the present setting because (*a*) signature development in high dimensional data requires large sample sizes, and (*b*) when the fraction of sensitive patients is low, a large number of patients needs to be screened to identify the sufficient number of sensitive patients to achieve acceptable power.

In this article, we describe an extension of the ASD in which signature development and validation are embedded in a complete cross-validation procedure. This allows the use of virtually the entire study population in both signature development and validation steps. We develop a procedure that preserves the study-wise type I error while substantially increasing the statistical power for establishing a statistically significant treatment effect for an identified subset of patients who benefit from the experimental treatment. We also examine approaches to estimation of treatment effect for the identified sensitive subset.

Authors' Affiliations: [1]Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, Maryland and [2]Department of Mathematics and Statistics, Queen's University, Kingston, Ontario, Canada

Corresponding Author: Boris Freidlin, Biometric Research Branch, EPN-8122, National Cancer Institute, Bethesda, MD 20892. Phone: 301-402-0640; Fax: 301-402-0560; E-mail: freidlinb@ctep.nci.nih.gov.

doi: 10.1158/1078-0432.CCR-09-1357

Cross-Validated Adaptive Signature Design End of Trial Analysis

- Compare T to C for **all patients** at significance level $\alpha_{overall}$
  - If overall H_0 is rejected, then claim effectiveness of T for eligible patients
  - Otherwise

- Using a pre-specified classifier development algorithm $\mathcal{A}$, develop a predictive binary classifier C on the full dataset
  - This may involve variable selection and/or tuning parameter optimization
  - C is a binary classifier with C(x)=1 means patient is predicted to benefit from E over C

Resubstitution estimate of treatment effect

- $S=\{i \mid C(x_i)=1\}$

- T=estimated treatment effect (e.g. log hazard ratio or log-rank statistic) in S

De-biasing T

- Let Δ_S=true treatment effect in S
- $T=\Delta_S$- bias
  - $\Delta_S = T + \text{bias} \approx T+\Delta_{resamp} - T$
- Δ_{resamp}= re-sampling estimate of $E_S[\Delta_S]$

Re-sampling estimate of $E_S[\Delta_S]$

- Partition the full data set into K parts
- Form a training set by omitting one of the K parts. The omitted part is the test set
  - Apply classifier development algorithm $\mathcal{A}$ to the training set, develop a predictive classifier C'of the subset of patients who benefit preferentially from the new treatment E
  - Classify the patients in the test set as likely to benefit from E or not
- Repeat this procedure K times, leaving out a different part each time
  - After this is completed, all patients in the full dataset are classified. Let S' denote the patients classified as sensitive to E

- Compare E to C in S', computing a measure of difference Δ_{resamp} . This might be the difference in response proportions or for survival data the log-hazard ratio or log-rank statistic

- Δ_{resamp} is the estimate of measure of treatment effect in patients who are selected for treatment by the classifier C developed by applying $\mathcal{A}$ to the full dataset.
- Generate the null distribution of Δ_{resamp} by permuting the treatment labels and repeating the entire K-fold cross-validation procedure
- Perform test at significance level 0.05 - $\alpha_{overall}$
- If H_0 is rejected, claim effectiveness of E for subset defined by classifier C

80% Response to T in Sensitive Patients
25% Response to C otherwise
25% Response to C
10% Patients Sensitive

	ASD	CV-ASD
Overall 0.05 Test	0.223	0.240
Overall 0.04 Test	0.198	0.209
Sensitive Subset 0.01 Test	0.205	0.661
Overall Power	0.351	0.714

70% Response to T in Sensitive Patients
25% Response to T Otherwise
25% Response to C
20% Patients Sensitive

	ASD	CV-ASD
Overall 0.05 Test	0.486	0.503
Overall 0.04 Test	0.452	0.471
Sensitive Subset 0.01 Test	0.207	0.588
Overall Power	0.525	0.731

70% Response to T in Sensitive Patients
25% Response to T Otherwise
25% Response to C
30% Patients Sensitive

	ASD	CV-ASD
Overall 0.05 Test	0.830	0.838
Overall 0.04 Test	0.794	0.808
Sensitive Subset 0.01 Test	0.306	0.723
Overall Power	0.825	0.918

35% Response to T
25% Response to C
No Subset Effect

	ASD	CV-ASD
Overall 0.05 Test	0.586	0.594
Overall 0.04 Test	0.546	0.554
Sensitive Subset 0.01 Test	0.009	0
Overall Power	0.546	0.554

25% Response to T
25% Response to C
No Subset Effect

	ASD	CV-ASD
Overall 0.05 Test	0.047	0.056
Overall 0.04 Test	0.04	0.048
Sensitive Subset 0.01 Test	0.001	0
Overall Power	0.041	0.048

506 prostate cancer patients were randomly allocated to one of four arms: Placebo and 0.2 mg of diethylstilbestrol (DES) were combined as control arm C

1.0 mg DES, or 5.0 mg DES were combined as E.

The end-point was overall survival (death from any cause).

Covariates:
Age: In years
Performance status (pf): Not bed-ridden at all vs other
Tumor size (sz): Size of the primary tumor (cm2)
Index of a combination of tumor stage and histologic grade (sg)
Serum phosphatic acid phosphatase levels (ap)

485 cases with all covariates

A proportional hazards regression model was developed using patients in both E and C groups. Main effect of treatment, main effect of covariates and treatment by covariate interactions were considered.

$\log[HR(z,x)] = a\,z + b'x + z\,c'x$

$z = 0,1$ treatment indicator ($z=0$ for control)

x = vector of covariates

$\log[HR(1,x)] - \log[HR(0,x)] = a + c'x$

Define classifier $C(X) = 1$ if $a + c'x < c$

$= 0$ otherwise

c = median of the $(a + c'x)$ values in the training set.

Figure 1: Overall analysis. The value of the log-rank statistic is 2.9 and the corresponding p-value is 0.09. The new treatment thus shows no benefit overall at the 0.05 level.

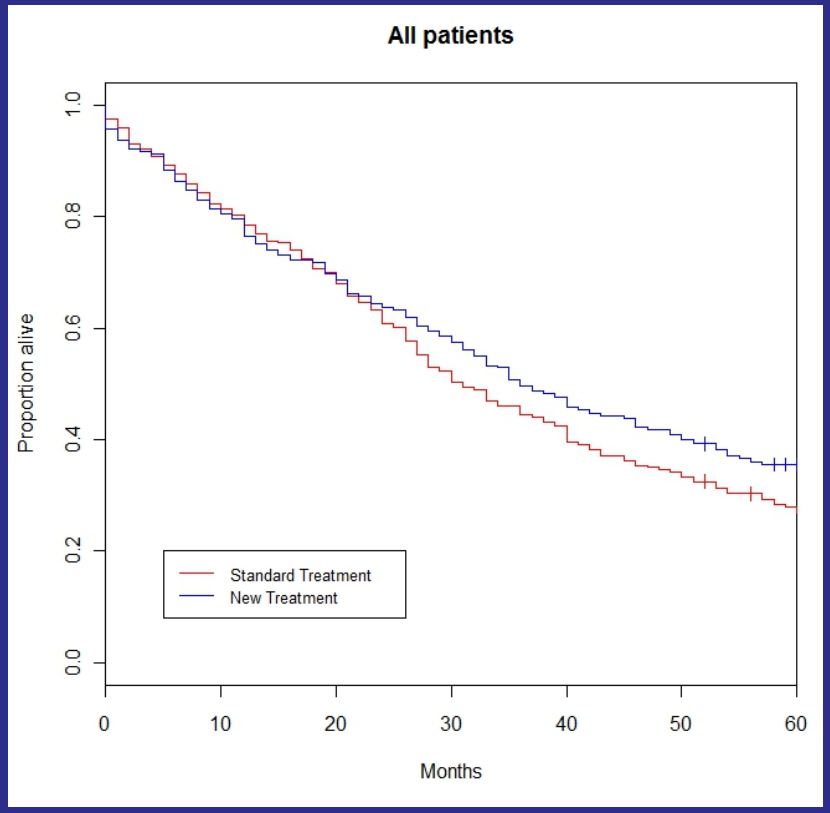


Figure 2: Cross-validated survival curves for patients predicted to benefit from the new treatment. log-rank statistic = 10.0, permutation p-value is .002

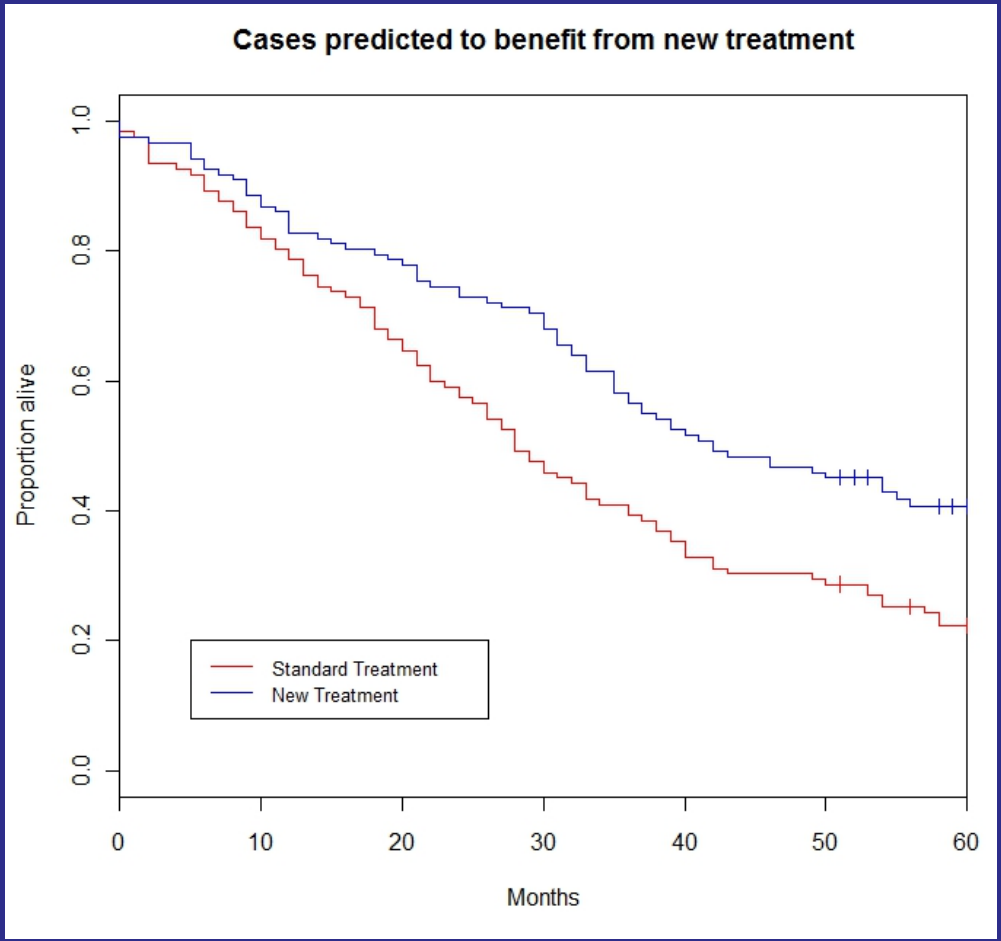
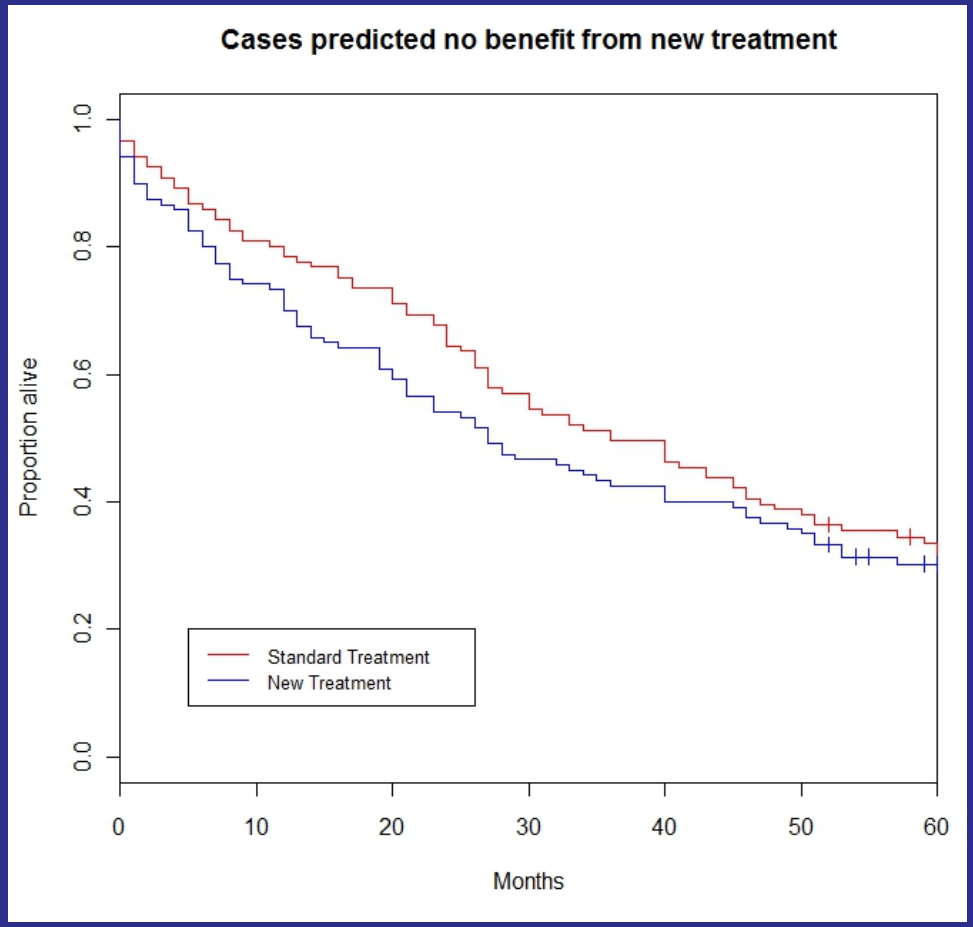


Figure 3: Survival curves for cases predicted not to benefit from the new treatment. The value of the log-rank statistic is 0.54.



Proportional Hazards Model Fitted to Full Dataset

	coef	p-value
Treatment	-2.195	0.12
age	0.002	0.85
pf(Normal.Activity)	-0.260	0.25
sz	0.020	0.001
sg	0.113	0.004
ap	0.002	0.21
Treatment*age	0.050	0.003
Treatment*pf(Normal.Activity)	-0.743	0.026
Treatment*sz	-0.010	0.26
Treatment*sg	-0.074	0.19
Treatment*ap	-0.003	0.11

Published OnlineFirst August 27, 2012; DOI: 10.1158/1078-0432.CCR-12-1206

Clinical Cancer Research

Predictive Biomarkers and Personalized Medicine

Developing and Validating Continuous Genomic Signatures in Randomized Clinical Trials for Predictive Medicine

Shigeyuki Matsui[1], Richard Simon[2], Pingping Qu[3], John D. Shaughnessy Jr[4], Bart Barlogie[4], and John Crowley[3]

Abstract

Purpose: It is highly challenging to develop reliable diagnostic tests to predict patients' responsiveness to anticancer treatments on clinical endpoints before commencing the definitive phase III randomized trial. Development and validation of genomic signatures in the randomized trial can be a promising solution. Such signatures are required to predict quantitatively the underlying heterogeneity in the magnitude of treatment effects.

Experimental Design: We propose a framework for developing and validating genomic signatures in randomized trials. Codevelopment of predictive and prognostic signatures can allow prediction of patient-level survival curves as basic diagnostic tools for treating individual patients.

Results: We applied our framework to gene-expression microarray data from a large-scale randomized trial to determine whether the addition of thalidomide improves survival for patients with multiple myeloma. The results indicated that approximately half of the patients were responsive to thalidomide, and the average improvement in survival for the responsive patients was statistically significant. Cross-validated patient-level survival curves were developed to predict survival distributions of individual future patients as a function of whether or not they are treated with thalidomide and with regard to their baseline prognostic and predictive signature indices.

Conclusion: The proposed framework represents an important step toward reliable predictive medicine. It provides an internally validated mechanism for using randomized clinical trials to assess treatment efficacy for a patient population in a manner that takes into consideration the heterogeneity in patients' responsiveness to treatment. It also provides cross-validated patient-level survival curves that can be used for selecting treatments for future patients. *Clin Cancer Res; 18(21); 6065–73.* ©2012 AACR.

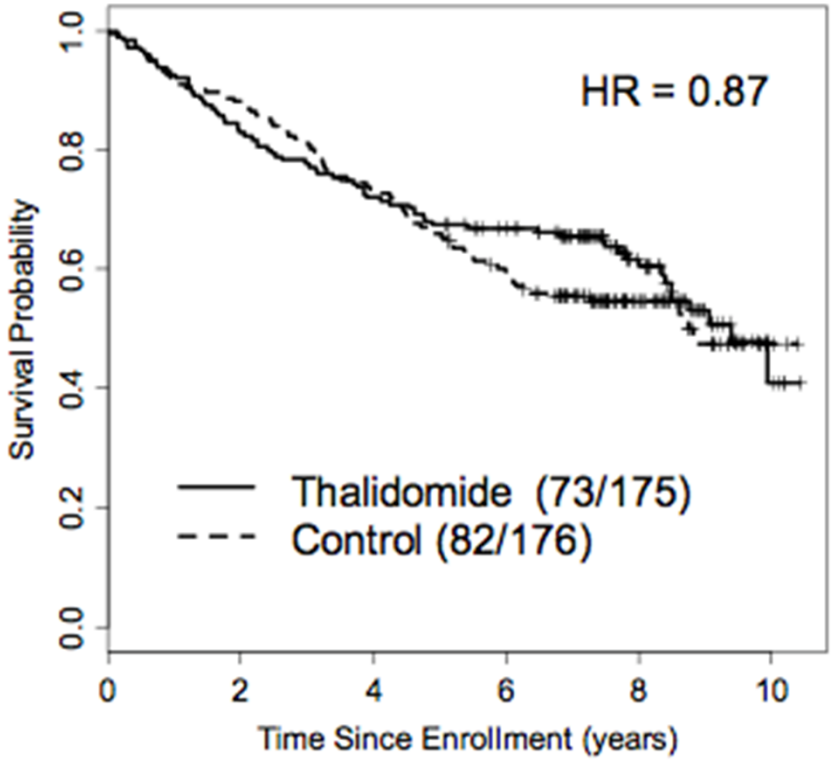


Fig. 1. Survival curves for all 351 patients with genomic data in the randomized trial for multiple myeloma.

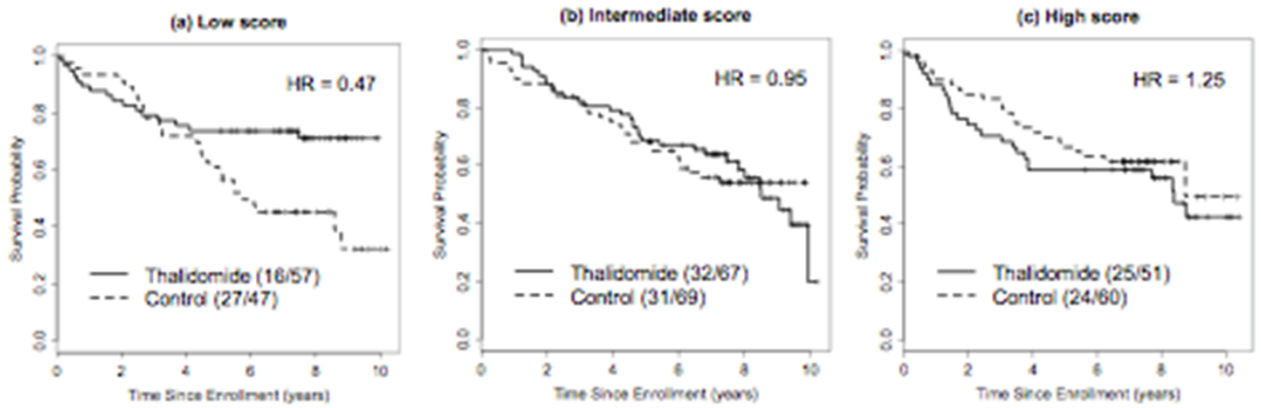


Fig. 5. Survival curves for each of the three subclasses, "Low", "Intermediate" and "High" derived from using thresholds of 33rd and 66th percentiles in the predicted signature score *S* (panels a-c).

Acknowledgements

- Adaptive randomization
  - David Hoel
  - George Weiss
- Adaptive stratification
  - Stuart Pocock
- Adaptive sample size
  - Gordon Lan
  - Max Halperin
- Adaptive rx selection
  - Peter Thall
  - Susan Ellenberg
- Enrichment designs
  - Aboubakar Maitournam
- Run-in design
  - Fangxin Hong
- Adaptive enrichment
  - Noah Simon
- Adaptive target population
  - Boris Freidlin
  - Wenyu Jiang
  - Shigeyuki Matsui