Bayesian Oncology Clinical Trial Designs with Subgroup-Specific Decisions

Peter F. Thall, PhD
Department of Biostatistics
The University of Texas
M.D. Anderson Cancer Center

2019 OncoStat Meeting
Precision Oncology Trials
April 26-27, 2019
Hartford, CN

Design 1: Randomized Comparison of Nutritional Prehabilitation (N) to Standard of Care (C) based on Post Operative Morbidity after Esophageal Resection

Collaborators: **Thomas Murray**, U. Minnesota

Ying Yuan, Joan Elizondo, Wayne Hofstsetter (PI), MD Anderson

Goal: Compare N to C allowing different conclusions in P = Primary (60%) and S = Salvage (40%) patients

Outcome: Clavien-Dindo postoperative morbidity (POM) score within days 30 post surgery (0= normal recovery, . . . , 5 = death)

N_{max} = **200 patients** (approx. 2 years, multi-institution). Stratified randomization in blocks of size four. Interim analysis at 100 patients, may stop early for superiority of either N or C.

Trial is ongoing per MDACC protocol 2017-0772. Statistical design paper: Murray, et. al. *Biometrics* 74:1095-1103, 2018

Reducing Post Operative Morbidity

Y = 30-day POM score has possible values $y = 0, 1, 2, 3, 4, 5$.

A robust Bayesian hierarchical regression model is assumed for

$$\Pr(Y = y \mid \textit{Treatment}, \textit{Subgroup})$$

- Treatment = Nuprehab or Control (N or C)
- Subgroup = Primary or Salvage (P or S)

The probability model
- Borrows strength between data from the two subgroups
- Allows treatment$\times$subgroup interactions

Model prior parameters were determined from information on POM score elicited from the trial PI.

Reducing Post Operative Morbidity

Elicited prior POM score Probabilities
for C= Standard of Care

	0	1	2	3	4	5
Primary	.50	.20	.10	.10	.05	.05
Salvage	.30	.25	.10	.10	.10	.15

Elicited numerical POM score **Utilities**

Score	0	1	2	3	4	5
Utility	100	85	65	25	10	0

Subgroup-Specific interim and final N-versus-C tests are based on $\Pr\{\overline{U}(N, g, \theta) > \overline{U}(C, g, \theta)\}$ where

$\overline{U}(N, g, \theta)$ = Mean Utility of N in subgroup $g = P$ or S

$\overline{U}(C, g, \theta)$ = Mean Utility of C in subgroup $g = P$ or S

Reducing Post Operative Morbidity

Operating Characteristics of the Subgroup-Specific Utility-Based Design, Nmax = 200, for 4 scenarios. Correct decision probabilities are given in **blue**.

	Pr Conclude N Superior to C		Pr Conclude N Inferior to C		Mean *N*
Scenario	Prim	Salv	Prim	Salv	
1 (Null/Null)	.02	.02	.03	.03	199.2
2 (**Alt**/Null)	**.78**	.04	**.00**	.02	189.6
3 (Null/**Alt**)	.03	**.80**	.02	**.00**	187.0
4 (**Alt/Alt**)	**.82**	**.84**	**.00**	**.00**	172.4

Reducing Post Operative Morbidity

Operating Characteristics of the **DUMBED DOWN "One Size Fits All" Design that IGNORES SUBGROUPS**. Correct decision probabilities are given in **blue**.

	Pr Conclude N Superior to C		Pr Conclude N Inferior to C		Mean N
Scen(Prim/Salv)	Prim	Salv	Prim	Salv	
1 (Null/Null)	.02	.02	.03	.03	199.4
2 (**Alt**/Null)	**.44**	.44	**.00**	.00	193.0
3 (Null/**Alt**)	.56	**.56**	.00	**.00**	189.6
4 (**Alt/Alt**)	**.98**	**.98**	**.00**	**.00**	145.1

In Scenario 2 : Pr(Type II error | Primary) $= 1 - .44 =$ **.56**, and Pr(Type I error | Salvage) $=$ **.44**. $\Rightarrow$ **Why not just flip a coin to decide which treatment is better?** If you are lucky and there is no treatment-subgroup effect (Scenario 4), the dumbed-down design has very high power.

Design # 2: Optimizing Natural Killer Cell Doses for Heterogeneous Cancer Patients Based on Multiple Event Times

Collaborators: Juhee Lee, UC Santa Cruz and Katy Rezvani (PI), MD Anderson

Natural killer (NK) cells are a subset of lymphocytes that can be used for cancer immunotherapy (Rezvani and Rouce, 2015).

A Bayesian sequentially adaptive design is presented for an early phase clinical trial to optimize subgroup-specific doses of umbilical cord blood derived NK cells as therapy for severe hematologic malignancies. **Trial ongoing, per MDACC protocol 2017-0295**.

Statistical paper to appear in *J Royal Statistical Society, Series C*

Medical Background

Treatments for severe hematologic malignancies:

Chemotherapy : Often does not work, with either no disease remission or cancer recurrence soon after remission.

Allogeneic Stem Cell Transplantation: Provides longer survival, on average, but carries risks of graft-versus-host disease, toxicity, infection, regimen-related death. Often used as a salvage therapy following failure of chemo.

Adoptive cell Therapy: A new therapeutic modality for advanced cancers refractory to conventional therapy. Grow specialized cells *ex vivo*, then infuse them into the patient.

- Chimeric antigen receptor (CAR) T-cells
- NK cells

A Trial of Natural Killer Cell Therapy

Patient Subgroups

1. **Type of B-cell Malignancy**:

CLL = Chronic Lymphocytic Leukemia
ALL = Acute Lymphocytic Leukemia
NHL = Non-Hodgkin's Lymphoma

2. **Disease Bulk**: Low (LBD) or High (HBD)

$\Rightarrow 3 \times 2 = 6$ (Disease Type, Disease Bulk) prognostic subgroups.

Primary Goals: Within each subgroup ("Precision Medicine")

1. Do interim safety monitoring of 100-day death rates

2. Identify optimal NK cell dose, from $\{10^5, 10^6, 10^7\}$ cells per kg

Clinical Outcomes

Conventional Phase I: Toxicity alone used to find "MTD"

Less often, (Efficacy, Toxicity) used together in phase I-II.

The NK cell trial has *Five Co-Primary Time-to-Event Outcomes:*

The times, from NK cell infusion, to D=Death, and the four nonfatal events P = Progressive Disease, R=Response, T=Severe Toxicity, and C=severe Cytokine Release Syndrome.

Monitor P, R, and D for 365 days.

T and C are most likely to occur soon after NK cell infusion $\Rightarrow$ monitor T and C for 100 days

Clinical Outcomes

Y_j = time to event $j = P, R, T, C, D$

1. Informative censoring of Y_P, Y_R, Y_T, Y_C by death
2. P and R are competing risks, since at most one can occur
3. Based on clinical experience, the outcomes are highly interdependent, with positive associations between the adverse event times.
4. The distribution of the event times $\boldsymbol{Y} = (Y_R, Y_C, Y_P, Y_T, Y_D)$ varies between the 6 subgroups.
5. These 5 event times all are Potential Outcomes: Each may or may not occur within its follow up period.

Monitoring and Logistics

Each event time is fully evaluated at

- its occurrence time, or
- the time of death, or
- the end of its follow up period if it has not occurred and the patient is alive

Main Logistical Problem:

It is not feasible to repeatedly suspend accrual to wait for full evaluation of all previously treated patients' outcomes to choose each new patient's NK cell dose adaptively.

Challenges

- Ignore subgroups $\Rightarrow$ High probabilities of making incorrect decisions within subgroups.
- Small-to-moderate sample sizes in early phase trials limit the reliability of subgroup-specific decisions.
- Based on current knowledge about NK cell biology, the rate of each outcome **may or may not increase with NK cell dose**.

	LBD (Z=0)	HBD (Z=1)
CLL (r=1)	7	13
ALL (r=2)	7	13
NHL (r=3)	7	13

- $Z = 0$ for LBD, $Z = 1$ for HBD
- $r = 1, 2, 3$ for CLL, ALL, NHL

$\Leftarrow$ Expected subsample sizes with $N_{\max} = 60$ patients

Statistical Model

The assumed Bayesian statistical models for regression of each Y_j = time to event $j = P, R, T, C,$ or D depend on

- HBD-vs-LBD effect on the event rate
- NK cell dose effect on event rate for each disease type
- a vector of 5 random patient-specific latent frailties

The frailty vector is not observed. It is a Bayesian model component that

1. accounts for variability not explained by (NK cell dose, disease type, disease severity)

2. induces associations among the 5 potential event time outcomes.

Dose Evaluation - Optimality

Early follow up intervals 30 days for Y_C and 100 days for all other outcomes $\Rightarrow$

$3 \times 2^3 = 24$ possible early outcomes $\boldsymbol{\delta} = (\delta_C, \delta_R, \delta_T, \delta_P, \delta_D) =$ indicators of occurrence or not during the early follow up periods

Utility Elicitation

1. We first set minimum utility $U(\boldsymbol{\delta}) = 0$ if $\delta_D = 1$ (DEATH within 100 days)

2. We then elicited numerical utilities for each of the 12 events for patients who survived 100 days

Outcome Utilities

Elicited utilities of outcomes for patients alive at day 100

		(δ_P, δ_R)		
δ_C	δ_T	(1,0)	(0,0)	(0,1)
0	0	**20**	**50**	**90**
0	1	**10**	**30**	**70**
1	0	**10**	**30**	**70**
1	1	**5**	**20**	**50**

Dose Optimality for each $\boldsymbol{Z}$ is based on the joint probability distribution of $\boldsymbol{\delta}$ given the observed data

Dose Safety for each $\boldsymbol{Z}$ is based on the probability distribution of δ_D given the observed data

Trial Conduct

For each NK cell dose d and disease subgroup $\boldsymbol{Z}$, denote

$$\pi_D(d, \boldsymbol{Z}, \boldsymbol{\theta}) = \textbf{Probability of Death within 100 days}$$

During the trial, for upper limit $\bar{\pi}_D(\boldsymbol{Z})$ specified by the PI, if

$$\Pr\{\pi_D(d, \boldsymbol{Z}, \boldsymbol{\theta}) > \bar{\pi}_D(\boldsymbol{Z}) \mid \textit{the observed data}\} > .80$$

then d is considered unsafe for subgroup $\boldsymbol{Z}$, and is no longer administered to patients in that subgroup.

For the 6 subgroups, elicited $\bar{\pi}_D(\boldsymbol{Z})$ values varied from .15 to .40 .

Trial Conduct

- Safety monitoring is begun for each disease type when 9 patients have been enrolled and at least 5 of the 9 have died or been followed for 100 days.

- For each Z, decide which doses are safe or unsafe. Do not administer unsafe doses. If no dose is safe for that subgroup, stop enrolling patients in that subgroup.

- As data are accumulated, the set of "safe" doses for each Z may change adaptively. So, a dose may go from safe to unsafe, but later go back to being considered safe.

Additional Safety Rule

Additional rule imposed by a **Federal Regulatory Agency**:

1. Ignore disease subgroups.
2. Assume Pr(Death within 30 days | d) increases in d.
3. Stop the trial if the probability of death within 30 days at the lowest NK cell dose is "too high."

To do this, we formulated a **Dumbed-Down Model** and a **One-Size-Fits-All FRA Safety Stopping Rule** based on the probability of death within 30 days, ignoring subgroups

The Bad News: The FRA Rule may stop the trial for all patients, regardless of disease subgroup

Trial Conduct

Z = (disease severity, disease type). During the trial, for each disease type ALL, CLL, NHL, **randomize patients among the three NK cell doses in order of entry to the trial by randomly permuting the integers (1,2,3)**.

For each patient :

1. record date and dose of NK cells administered,
2. repeatedly update all 5 event times, Y_C, Y_T, Y_R, Y_P, Y_D, by recording (i) the occurrence date or (ii) the last follow up date without occurrence, **exactly as is done with survival data**, so each adaptive decision is made based on the most current data.

Final Trial Conclusions

Determine a final optimal dose for each $\boldsymbol{Z}$:

1. If no dose is safe $\Rightarrow$ No dose is selected
2. If at least one dose is safe, then select the dose having highest mean utility given the final data

Computer Simulation Study of the Design

Evaluation Criteria for a given Subgroup Z

P_{stop} = Probability of identifying a truly unsafe dose with probability of death within 100 days larger than the fixed limit $\overline{\pi}_D(Z)$

P_{sel} = Probability of selecting the optimal safe dose for Z

Simulation: Scenario 2

Using patient subgroup information is critical.

prognostic subgroup ($\boldsymbol{Z}$)

	LBD	HBD
CLL	(0, 1)	(1, 1)
ALL	(0, 2)	(1, 2)
NHL	(0, 3)	(1, 3)

Dose	$d=1$	$d=2$	$d=3$	$\bar{\pi}_D$	$d=1$	$d=2$	$d=3$	$\bar{\pi}_D$
π_D^{TR}	0.02	0.45	0.60	*0.15*	0.04	0.70	0.84	*0.30*
$\bar{U}^{\mathrm{TR}}$	42.34	22.86	16.00		39.73	9.49	4.63	
$\mathrm{P}_{\mathrm{stop}}$	0.00	0.89	0.97		0.00	0.98	1.00	
$\mathrm{P}_{\mathrm{sel}}$	1.00	0.00	0.00		1.00	0.00	0.00	
π_D^{TR}	0.40	0.60	0.05	*0.20*	0.64	0.84	0.10	*0.40*
$\bar{U}^{\mathrm{TR}}$	23.81	15.82	40.41		10.99	4.62	36.02	
$\mathrm{P}_{\mathrm{stop}}$	0.67	0.92	0.00		0.81	0.98	0.00	
$\mathrm{P}_{\mathrm{sel}}$	0.00	0.00	1.00		0.00	0.00	1.00	
π_D^{TR}	0.65	0.05	0.35	*0.20*	0.88	0.10	0.58	*0.40*
$\bar{U}^{\mathrm{TR}}$	14.03	40.47	26.31		3.46	36.08	13.26	
$\mathrm{P}_{\mathrm{stop}}$	0.95	0.00	0.55		1.00	0.00	0.68	
$\mathrm{P}_{\mathrm{sel}}$	0.00	1.00	0.00		0.00	1.00	0.00	

Simulation: Scenario 3

All doses are unsafe for all $\boldsymbol{Z}$.

prognostic subgroup ($\boldsymbol{Z}$)	LBD	HBD
CLL	(0, 1)	(1, 1)
ALL	(0, 2)	(1, 2)
NHL	(0, 3)	(1, 3)

Dose	$d=1$	$d=2$	$d=3$	$\bar{\pi}_D$	$d=1$	$d=2$	$d=3$	$\bar{\pi}_D$
π_D^{TR}	0.42	0.38	0.37	*0.15*	0.66	0.62	0.60	*0.30*
$\bar{U}^{\mathrm{TR}}$	40.33	44.40	<u>44.55</u>		20.71	24.52	<u>24.81</u>	
$\mathrm{P}_{\mathrm{stop}}$	0.88	0.85	0.82		0.96	0.95	0.94	
$\mathrm{P}_{\mathrm{sel}}$	0.07	0.11	0.13		0.03	0.05	0.05	
π_D^{TR}	0.52	0.58	0.65	*0.20*	0.77	0.83	0.88	*0.40*
$\bar{U}^{\mathrm{TR}}$	33.99	29.43	24.52		14.34	11.01	7.57	
$\mathrm{P}_{\mathrm{stop}}$	0.93	0.96	0.98		0.99	0.99	1.00	
$\mathrm{P}_{\mathrm{sel}}$	0.06	0.03	0.01		0.01	0.01	0.00	
π_D^{TR}	0.40	0.42	0.45	*0.20*	0.64	0.67	0.70	*0.40*
$\bar{U}^{\mathrm{TR}}$	42.49	40.21	38.79		22.61	20.32	18.95	
$\mathrm{P}_{\mathrm{stop}}$	0.73	0.77	0.84		0.85	0.87	0.94	
$\mathrm{P}_{\mathrm{sel}}$	0.19	0.16	0.07		0.12	0.11	0.04	

Simulation: Scenario 6

$\bar{U}^{\text{TR}}$ varies with (d, Z, r), and the set of acceptable doses varies with $\boldsymbol{Z} = (Z, r)$.

prognostic subgroup ($\boldsymbol{Z}$)

	LBD	HBD
CLL	(0, 1)	(1, 1)
ALL	(0, 2)	(1, 2)
NHL	(0, 3)	(1, 3)

Dose	$d=1$	$d=2$	$d=3$	$\bar{\pi}_D$	$d=1$	$d=2$	$d=3$	$\bar{\pi}_D$
π_D^{TR}	0.35	0.03	0.13	*0.15*	0.75	0.10	0.37	*0.30*
$\bar{U}^{\text{TR}}$	41.74	59.80	57.69		14.53	55.48	40.90	
P_{stop}	0.76	0.00	0.09		0.99	0.00	0.34	
P_{sel}	0.00	0.56	0.44		0.00	0.99	0.01	
π_D^{TR}	0.08	0.45	0.02	*0.20*	0.24	0.86	0.06	*0.40*
$\bar{U}^{\text{TR}}$	57.75	34.95	57.83		47.19	7.81	55.07	
P_{stop}	0.00	0.82	0.00		0.02	0.99	0.00	
P_{sel}	0.80	0.00	0.20		0.04	0.00	0.96	
π_D^{TR}	0.05	0.10	0.30	*0.20*	0.16	0.29	0.70	*0.40*
$\bar{U}^{\text{TR}}$	59.50	57.18	46.28		52.57	44.10	18.84	
P_{stop}	0.00	0.01	0.51		0.00	0.03	0.91	
P_{sel}	0.53	0.47	0.01		0.89	0.11	0.00	

Simulation Summary

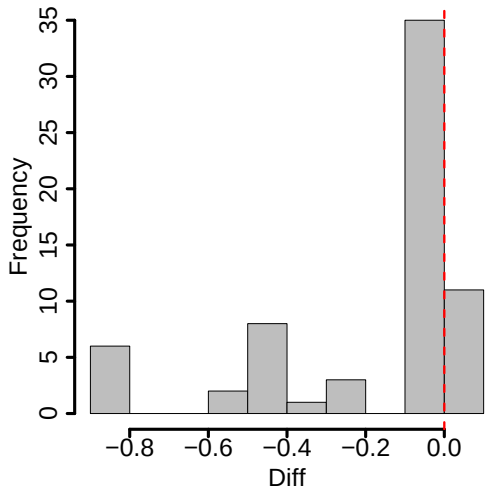
Reliability The design captures the pattern of $\bar{U}^{\mathrm{TR}}$ as a function of (dose, subgroup) quite well and makes correct decisions with high probabilities.

Safety The design reliably identifies unsafe doses for each subgroup and assigns fewer patients to doses declared unsafe.

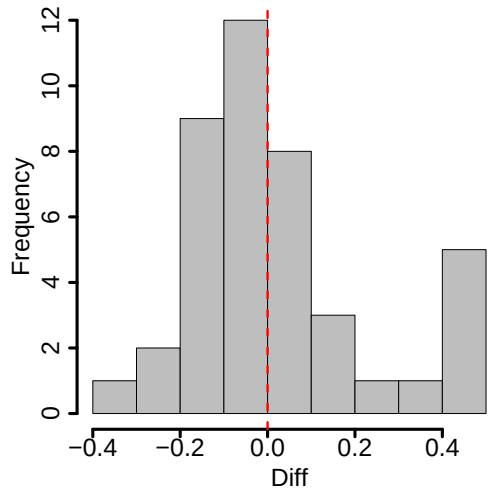
Protection Against Stupidity The FRA rule rarely terminates the trial because the design's rules supersede it. E.g. if only d=1 is unsafe in one or two subgroups, the design terminates accrual to d=1 in those subgroups.
This stops the FRA rule from incorrectly terminating the entire trial.

Gains from Being Precise

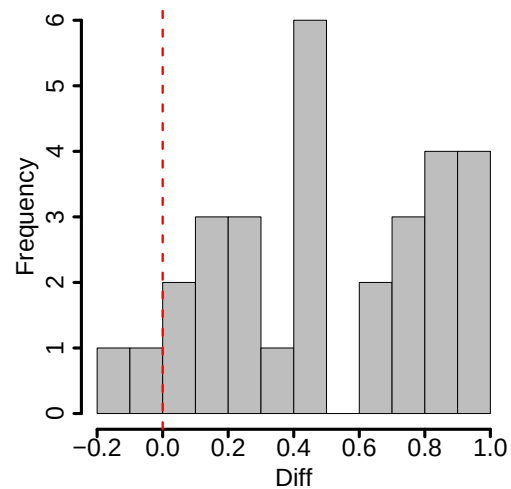
Comparison to a dumbed down version of the design that ignores the disease subgroup $\boldsymbol{Z}$ and makes "one size fits all" decisions.



(a) $\mathsf{P}_{\mathsf{stop}}(d, \boldsymbol{Z}) - \mathsf{P}_{\mathsf{stop}}(d)$ for truly safe doses



(b) $\mathsf{P}_{\mathsf{stop}}(d, \boldsymbol{Z}) - \mathsf{P}_{\mathsf{stop}}(d)$ for truly unsafe doses



(c) $\mathsf{P}_{\mathsf{sel}}(d, \boldsymbol{Z}) - \mathsf{P}_{\mathsf{sel}}(d)$ for truly optimal doses

Additional Simulations

We examined the design's robustness by simulating the event time outcomes from various distributions (log-logistic, log-normal)
Differences in the patterns of the hazard functions over time affect the design's performance very slightly $\Rightarrow$ The design is extremely robust.

- Increase sample size to $N_{\max} = 120$ under all scenarios $\Longrightarrow$
  - This greatly improves P_{stop} and P_{sel} for many $(d, \boldsymbol{Z}) \Longrightarrow$
  - **Larger $N_{\max}$ is highly desirable for "early phase" designs making subgroup-specific decisions**.

- If we make follow up shorter by using $\boldsymbol{L} = (100, 100, 100, 30, 100)$ in place of $\boldsymbol{L} = (365, 365, 100, 100, 365)$ then the design's performance deteriorates greatly .

General Conclusions

1. **Making subgroup-specific decisions is feasible** for early phase dose-finding trials or randomized comparative trials.

2. **Decades of clinical trials that have ignored patient heterogeneity probably have been a disaster**.

3. **Using elicited utilities of multiple outcomes provides a practical, ethical basis for decision making and treatment evaluation in clinical trials**.

This research was supported by NSF grant DMS-1662427 and NCI grant RO1 CA 83932.