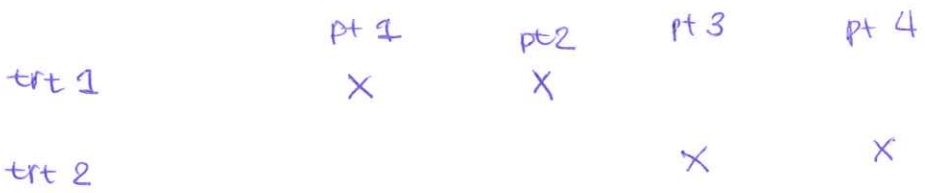
Treatment- and response-adaptive randomization procedures: probabilities of treatment assignment may change throughout the allocation process.

- Treatment-adaptive randomization procedures: assignment probabilities are adapted based on treatment imbalance.

  e.g. biased coin randomization and the urn design

- Response-adaptive randomization procedures: assignment probabilities are adapted based on observed responses or outcomes (will be discussed more later).

  e.g. play-the-winner rule



e.g. play-the-winner rule

- When a trial is used to compare two treatments and one of the treatments is superior to the other, many participants are given an inferior treatment for the duration of the trial.
- Attempt to allocate the better treatment to a great number of individuals.
- e.g. Zelen (1969): Consider a case where the outcome is binary, "success" or "failure"

  $\star\star$ The first participant is assigned one of the two treatments, with equal probability given to each.

  $\star\star$ If a success is observed, assign the following participant to the same treatment. If not, the other treatment is assigned.

* Example:

Adaptive Randomized Study of Idarubicin and Cytarabine Versus Troxacitabine and Cytarabine Versus Troxacitabine and Idarubicin in Untreated Patients 50 Years or Older With Adverse Karyotype Acute Myeloid Leukemia

By Francis J. Giles, Hagop M. Kantarjian, Jorge E. Cortes, Guillermo Garcia-Manero, Srdan Verstovsek, Stefan Faderl, Deborah A. Thomas, Alessandra Ferrajoli, Susan O'Brien, Jay K. Wathen, Lian-Chun Xiao, Donald A. Berry, and Elihu H. Estey

Purpose: **Troxacitabine has activity in refractory myeloid leukemia, either as a single agent or when combined with cytarabine (ara-C) or with idarubicin. A prospective, randomized study was conducted in patients aged 50 years or older with untreated, adverse karyotype, acute myeloid leukemia (AML) to assess troxacitabine-based regimes as induction therapy.**

Patients and Methods: **Patients were randomized to receive idarubicin and ara-C (IA) versus troxacitabine and ara-C (TA) versus troxacitabine and idarubicin (TI). A Bayesian design was used to adaptively randomly assign patients to treatment. Thus, although there was initially an equal chance for randomization to IA, TA, or TI, treatment arms with a higher success rate progressively received a greater proportion of patients.**

Results: **Thirty-four patients were treated. Randomization to TI stopped after five patients and randomization to TA stopped after 11 patients. Defining success as complete remission (CR) that occurred within 49 days of starting treatment, success rates were 55% (10 of 18 patients) with IA, 27% (three of 11 patients) with TA, and 0% (zero of five patients) with TI. Because three CRs occurred after day 49, final CR rates were 55% (10 of 18 patients) with IA, 45% (five of 11 patients) with TA, and 20% (one of five patients) with TI. The probability that TA was inferior to IA was 70%, with a 5% probability that TA would have a 20% higher CR rate than IA. Survival was equivalent with all three regimens.**

Conclusion: **Neither troxacitabine combination was superior to IA in elderly patients with previously untreated adverse karyotype AML.**

J Clin Oncol 21:1722-1727. © 2003 by American Society of Clinical Oncology.

- ~~IA~~ idarubicin + ara-C (IA) 18 patients X
- Troxacitabine + ara-C (TA) 11 patients
- " + idarubicin (TI) 5 patients

34

success: complete remission (CR)

prob of CR within 49 days

Data Monitoring and Interim Analysis (DM Chapter 10)

- *Interim Analysis*: Statistical analysis conducted during the course of a clinical trial for the purpose of monitoring efficacy and safety.

  $\Leftrightarrow$ *the repeated examination of evolving data*

- From the NIH policy, "a system for the appropriate oversight and monitoring of the conduct of clinical trials to ensure the safety of participants and the validity and integrity of the data for all NIH-supported or conducted clinical trials" is required.

- "The establishment of the data and safety monitoring boards (DSMBs) or data monitoring committees (DMCs) for multi-site clinical trials involving interventions that entail potential risk to the participants" is required.

- Determine if the trial should stop before its planned termination time. Possible reasons are
  - the superiority of the intervention under study is clearly established.

    since there is an ethical obligation not to continue to expose subjects to an inferior therapy.

  - if unacceptable adverse effects are apparent.

    since there is an ethical obligation to the study participants that a trial not continue beyond the point at which the potential risk to study participants outweigh the potential benefits.

* Example:

Significantly Higher Pathologic Complete Remission Rate After Neoadjuvant Therapy With Trastuzumab, Paclitaxel, and Epirubicin Chemotherapy: Results of a Randomized Trial in Human Epidermal Growth Factor Receptor 2–Positive Operable Breast Cancer

Aman U. Buzdar, Nuhad K. Ibrahim, Deborah Francis, Daniel J. Booser, Eva S. Thomas, Richard L. Theriault, Lajos Pusztai, Marjorie C. Green, Banu K. Arun, Sharon H. Giordano, Massimo Cristofanilli, Debra K. Frye, Terry L. Smith, Kelly K. Hunt, Sonja E. Singletary, Aysegul A. Sahin, Michael S. Ewer, Thomas A. Buchholz, Donald Berry, and Gabriel N. Hortobagyi

A B S T R A C T

Purpose

The objective of this study was to determine whether the addition of trastuzumab to chemotherapy in the neoadjuvant setting could increase pathologic complete response (pCR) rate in patients with human epidermal growth factor receptor 2 (HER2) –positive disease.

Patients and Methods

Forty-two patients with HER2-positive disease with operable breast cancer were randomly assigned to either four cycles of paclitaxel followed by four cycles of fluorouracil, epirubicin, and cyclophosphamide or to the same chemotherapy with simultaneous weekly trastuzumab for 24 weeks. The primary objective was to demonstrate a 20% improvement in pCR (assumed 21% to 41%) with the addition of trastuzumab to chemotherapy. The planned sample size was 164 patients.

* Example: (contd)

$p(y^* \mid y)$

Results

Prognostic factors were similar in the two groups. After 34 patients had completed therapy, the trial's Data Monitoring Committee stopped the trial because of superiority of trastuzumab plus chemotherapy. pCR rates were 25% and 66.7% for chemotherapy (n = 16) and trastuzumab plus chemotherapy (n = 18), respectively ($P = .02$). The decision was based on the calculation that, if study continued to 164 patients, there was a 95% probability that trastuzumab plus chemotherapy would be superior. Of the 42 randomized patients, 26% in the chemotherapy arm achieved pCR compared with 65.2% in the trastuzumab plus chemotherapy arm ($P = .016$). The safety of this approach is not established, although no clinical congestive heart failure was observed. A more than 10% decrease in the cardiac ejection fraction was observed in five and seven patients in the chemotherapy and trastuzumab plus chemotherapy arms, respectively.

Conclusion

Despite the small sample size, these data indicate that adding trastuzumab to chemotherapy, as used in this trial, significantly increased pCR without clinical congestive heart failure.

J Clin Oncol 23:3676-3685. © 2005 by American Society of Clinical Oncology

- **Sampling to a foregone conclusion**(DC 10.3, BCLM 2.5.4) The repeated examination of evolving data causes a problem.
- Reason: Since frequentists make decisions based on $p-$values (i.e., Type I error levels) and other design-based summaries, if we decide to look at the accumulating data many times over the course of the study, we must account for this in the procedure or risk inflating this Type I error.
- When the number of observations depends on the observations themselves, a fixed-sample-size analysis of the observations is **wrong**.
- There are some methods for performing interim analyses without an accompanying increase in the overall type I error rate.

e.g. [Anscombe, 1954] The mean μ of a normal population is to be estimated. Suppose the variance σ^2 is known.

* A first sample of n_1 (fixed) observations is taken and its mean $\bar{x}_1$ is calculated.
* If $\bar{x}_1$ does not differ significantly at the 5% level from some critical value, say 0, a second sample of n_2 observations is taken. Otherwise, no further observations are taken (n_2 is fixed).
  Let n_2 be much larger than n_1.
* Then if μ is in fact 0, the probability that the mean of all the observations will differ significantly from 0 at the 5% level, according to the ordinary fixed-sample-size test, is not 5% but 9.75%.
* This is because the sampling distribution of a function of observations is liable to be affected by the sequential sampling rule.

$\bar{x}_1$ { case 1: $\bar{x}_1$ is not sig. diff. from 0 $\Rightarrow$ n_2 more obs
case 2: $\bar{x}_1$ is sig. diff from 0 $\Rightarrow$ STOP

$\alpha = 0.05$

$P($ reject $H_0: \mu = 0 \mid$ true $H_0)$
$= 0.0975$

$N = \begin{cases} n_1 + n_2 \\ n_1 \end{cases}$ $\quad \bar{x} = \bar{x}_1$

$\frac{n_1\bar{x}_1 + n_2\bar{x}_2}{N}$

- Does this problem sound unfamiliar? No!
- From Anscombe (1954)

  "All risk of error is avoided if the method of analysis uses the observations only in the form of their likelihood function, since the likelihood function (given the observations) is independent of the sampling rule."

- From Statistical Decision Theory and Bayesian Analysis by Berger,

  The likelihood Principle: In making inferences or decisions about θ after x is observed, all relevant experimental information is contained in the likelihood function for the observed x. Furthermore, two likelihood functions contain the same information about θ if they are proportional to each other (as functions of θ)

T H T T H θ: prob of heads

$L(\theta|y) \propto \theta^2 (1-\theta)^3$

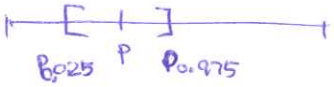
- In Bayesian statistics, decisions are based on posterior summaries that do not depend on how the experiment was stopped; the posterior simply evolves (typically narrowing) as data accumulate.
- Stopping for futility (stochastic curtailment–Frequentist): early termination on the grounds that the probability that a trial will be "successful" if allowed to continue is too small to justify continuation.
- Basic Idea: We compute the probability that a treatment will ever emerge as superior given the patients recruitment outlook and the data accumulated so far (a.k.a, predictive probability!).

  $\Rightarrow$ Then decision making: If this probability is too small, stop the trial.

- (BCLM Example 2.8) Suppose a medical device company wishes to run a safety study on one of its new cardiac pacemakers. Specifically, the company wishes to show that men receiving its new product will be very likely to be free from adverse events (AEs) during three months immediately following implantation of the device. Letting p be the probability a patient does not experience an AE in the first three months, we seek a 95% equal-tail Bayesian confidence interval for p, $(p_{.025}, p_{.975})$. Suppose our trial protocol uses the following decision rule:

  $$\text{Device is safe from AEs at 3 months} \Leftrightarrow p_{.025} > 0.85.$$

  That is, if the lower confidence bound for the chance of freedom from AEs is at least 85%, the trial succeeds; otherwise it fails.



- Suppose that we already have a preliminary study, Study A. We have $X_1 = 110$ and $n_1 = 117$. Our task is now to evaluate whether it is worth running a second dusty, Study B, which would enroll an additional n_2 patients.
- Suppose we begin with $U(0,1) = \text{Be}(1,1)$ prior for p.

Before Study A : $p \sim Be(1,1)$ $\quad X_{1i} \sim Bin(p)$ $\quad i=1,\dots,n_1$

After Study A : $p \mid x_1, \dots \sim Be(111, 8)$

$X_2 \mid p \sim Bin(n_2, p)$ & $p \sim Be(111, 8)$

- Evaluate $P(p_{0.025} > 0.85)$

Random $\nwarrow$ $p \in [p_{0.025}, p_{0.975}]$ after Study B

j : iteration

- Simulate $p_j \sim Be(111, 8)$
- Simulate $x_{2j} \sim Bin(n_2, p_j)$
- ~~Compute~~ update the post. of p using x_{2j} and compute $p_{0.025,j}$

$Be(111 + x_{2j},\ 8 + n_2 - x_{2j})$

$$\hat{P}(p_{0.025} > 0.85) = \frac{\#\text{ of } p_{0.025,j} > 0.85}{\#\text{ of iterations}}$$

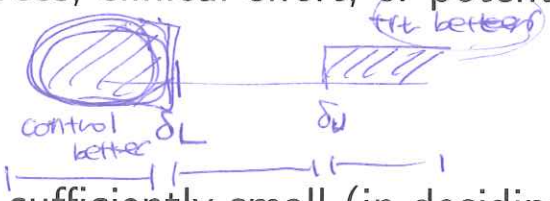
- Bayesian indifference zone methods (BCLM 2.5.2)
  - ⋆⋆ Δ: a treatment effect parameter

    positive ⇒ the trt is better
    negative ⇒ the control is better
  - ⋆⋆ Indifference zone (or range of equivalence) for a treatment effect parameter, $[\delta_L, \delta_U]$: a range of null Δ's over which we are indifferent between the intervention and the control.
    - ∗ δ_U (upper bound): the amount of improvement required by the intervention to suggest clinical superiority over control
    - ∗ δ_L (lower bound): the threshold below which the intervention would be considered clinically inferior.

ex1 0 — K (>0) additional benefit is required

ex2 $-\xi$ — $+\xi$

e.g. Suppose positive values of Δ are indicative of an efficacious treatment. Then set $\delta_U = K > 0$ and $\delta_L = 0$, an additional benefit perhaps being required of the treatment in order to justify its higher cost in terms of resources, clinical effort, or potential toxicity.



* May terminate the trial

(i) when $P(\Delta > \delta_U \mid \text{data})$ is sufficiently small (in deciding in favor of the control), *or*

(ii) when $P(\Delta < \delta_L \mid \text{data})$ is sufficiently small (in deciding in favor of the treatment)

Or

* Stop when one region's posterior probability is sufficiently large, or, failing this when a predetermined total sample size is reached.

Or

* Trial stopping rules might be based on the location of the 95% posterior credible interval for Δ, (Δ_L, Δ_U), with respect to the indifference zone $[\delta_L, \delta_U]$.

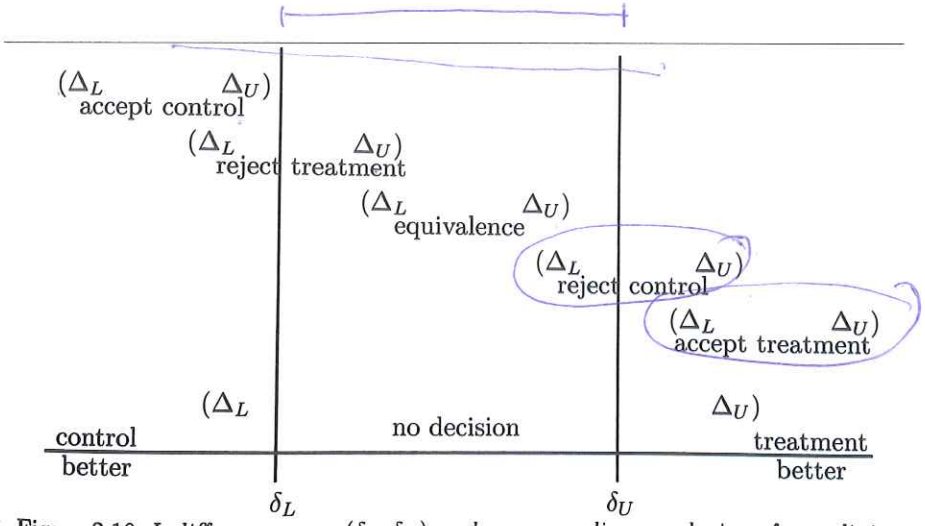


Figure 2.10 *Indifference zone (δ_L, δ_U) and corresponding conclusions for a clinical trial based on the location of the 95% posterior credible interval for Δ.*

An important feature in clinical trial design is **sample size** (DC Chapter 4).

"how big a sample size do I need for this trial?"

- If a clinical study is too small to yield meaningful results, it wastes resources and puts patients at risk with minimal potential for benefit.
- If a clinical study is too large, it may be able to identify effects that are too small to be of clinical interest while potentially exposing more subjects than necessary to inferior treatments.
- A maximum sample size is often given based on available resources
- The detectable effect sizes are determined from the sample size.

- Frequentist operating characteristics, Type I error, Type II error, power are very important to the FDA and other regulators.
- **Goal:** Find the smallest sample size, N, so that the type I error rate is at most α and the type II error rate is at most β for a given alternative hypothesis.
- The power of the study is the rejection probability, $1-\beta$, under the alternative.
- Subjects may fail to comply with study procedures e.g. loss to follow-up, non-adherence.

  $\Rightarrow$ this affects the required sample size.
- See DC 4.2 for examples of sample size calculation for frequentist designs.

- Bayesians care about frequentist operating characteristic (long-run behavior of the procedures) as well.

  Reasoning Well-designed Bayesian trials also have excellent frequentist properties.

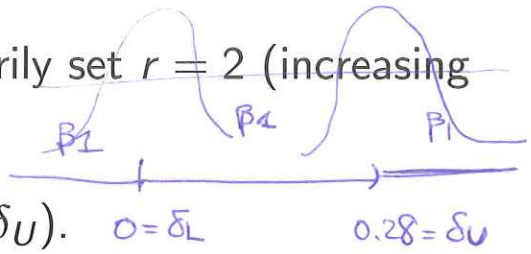
- **How?** We do *preposterior analysis*.

  ↺ One that averaged over the variability in both the unknown parameters and the as-yet unobserved data.

  ↺ Adding in the averaging over the prior leads to obvious preposterior Bayesian analogs of Type I error and power.

* **BCLM Example 2.9** (simulating power and Type I error for a Weibull survival model)

- t_i: the time until death for subject i in a clinical trial, with corresponding indicator x_i ($x_i = 0$ for control and 1 for treatment)
- Suppose $t_i \sim \text{Weibull}(r, \mu_i)$ where $r > 0$ and $\log(\mu_i) = -\beta_0 - \beta_1 x_i$.
- For the purpose of illustration, arbitrarily set $r = 2$ (increasing hazard over time)



- Now specify an indifference zone (δ_L, δ_U).
  - ⋆⋆ Say, we want to achieve a "clinically significant" improvement under the treatment $\Rightarrow \delta_L = 0$.
  - ⋆⋆ Say, we want to achieve a 15% increase in median survival $\Rightarrow \delta_U = 0.28$.

$S = \exp(-\mu_i t_i^r) = \exp(-\mu_i t_i^2)$

$= \exp(-e^{-\beta_0-\beta_1} t_i^2)$: trt.

$\exp(-e^{-\beta_0} t_{0i}^2)$: control

$\exp(-e^{-\beta_0-\beta_1} t_1^2) = 0.5$ $\quad t_1 = 1.15 t_0$

$\exp(-e^{-\beta_0} t_0^2) = 0.5$

* **BCLM Example 2.9** (contd)

- Now specify a prior – we employ "community of priors": skeptical, enthusiastic, and reference priors for the treatment effect parameter, β_1

  ⋆⋆ *Skeptical*: Let the prior mean equal to 0 (no change in survival) and choose the variance so that $P(\beta_1 > \delta_U) = \epsilon$ for a small but positive probability ϵ.

  e.g. $\epsilon = 0.05 \Rightarrow \beta_1 \sim \mathrm{N}(0, 0.17^2)$

  ⋆⋆ *Enthusiastic*: Let the prior mean to be δ_U but use the same standard deviation as the skeptical prior.

  ⋆⋆ *reference*: an improper uniform prior

* Priors

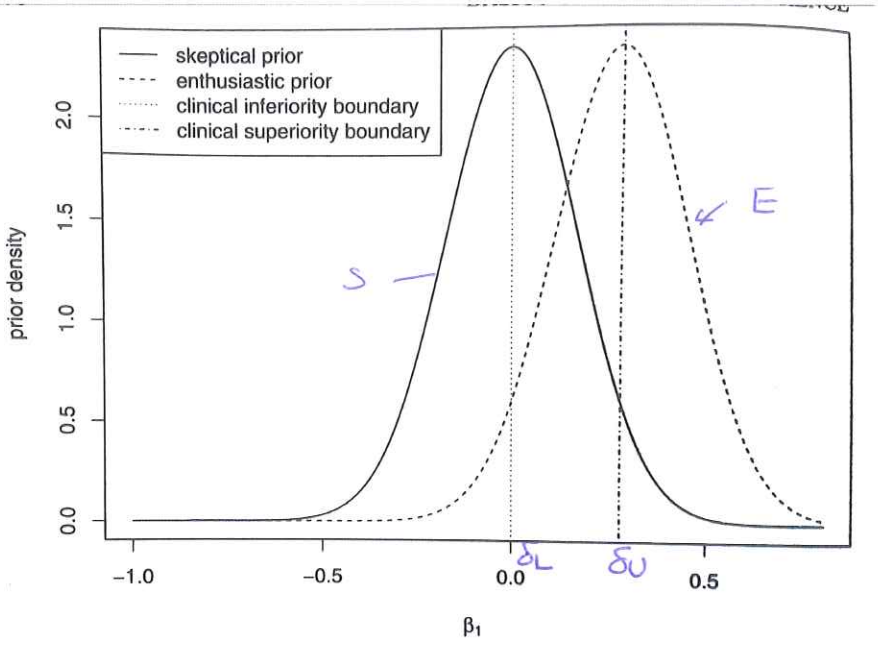


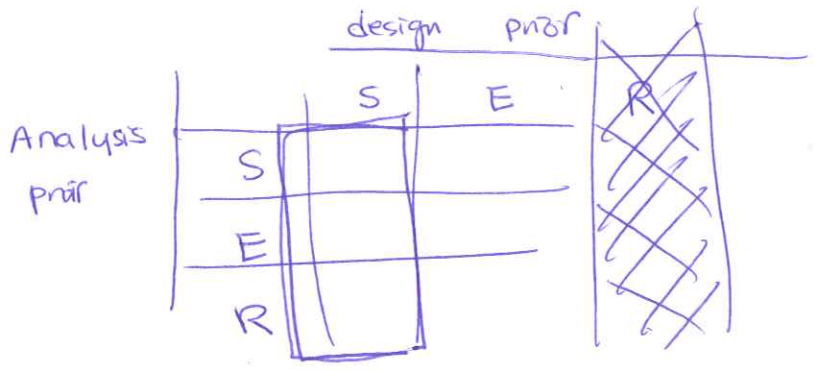
Figure 2.12 *Skeptical (solid line) and enthusiastic (dashed line) priors, Weibull survival model. Also shown (vertical lines) are the clinical inferiority boundary, $\delta_L = 0$, and the clinical superiority boundary, $\delta_U = 0.28$.*

* **BCLM Example 2.9** (contd)
  - Prior for β_0

    ** For β_0, let $\beta_0 \sim N(7.53, 0.2)$– somewhat informative and centered near values that emerge as plausible for our data generation mechanism

  - We use priors for two different things, a design prior and a analysis prior.

* **BCLM Example 2.9** (contd)
  - Suppose $N = 50$ for each group
  - Simulate β_0 and β_1 from a chosen prior (design prior)
  - Simulate t_i, $i = 1, \ldots, N$ for each group
  - Let's be a bit more realistic: Censoring.
    Simulate $c_i \overset{iid}{\sim} \mathrm{N}(80, 20^2) I(c > 0)$
  - Choose a analysis prior and compute the posterior credible interval of β_1
  - Make a decision compared to the indifference region.



* # of replications =100

```
Here are simulated outcome frequencies for N= 50
  accept control:     0
  reject treatment:   0.07
  equivalence:        0
  reject control:     0.87
  accept treatment:   0.06
  no decision:        0
End of BRugs power simulation
```

Bayesian Type II error

Bayesian power

* E : "truth"
* Use the ~~defined~~ R for analysis

control is better

Bayesian Type I error

* Prob of rejecting the control under a skeptical design prior

	analysis prior		
N	Skeptical	Reference	Enthusiastic
25	.001	.053	.178
50	.009	.069	.213
75	.017	.110	.209
100	.034	.070	.214

Table 2.4 *Probability of rejecting the control under a skeptical design prior for four sample sizes N and three analysis priors, Weibull survival model.*