Statistical Considerations for Early Phase and Biomarker Studies

Elizabeth G. Hill, PhD
Associate Professor of Biostatistics
Hollings Cancer Center
Medical University of South Carolina

Cancer Education Consortium
12 October 2013

Outline

- Overview
- Dose-finding trials
- Safety and efficacy trials
- Statistical and design considerations
- Correlative biomarker studies

Overview

Clinical trial definition

A *clinical trial* is an *experiment* testing a medical treatment on human subjects.

- Not all human research studies are *experimental*
- The *experiment* is what distinguishes clinical trials from other forms of medical research
- What do we mean by *experiment*?

"...the essential characteristic that distinguishes experimental from non-experimental studies is whether or not the scientist controls or manipulates the treatment ... under investigation." (Piantadosi, *Clinical Trials: A Methodologic Perspective*, 2005)

Equipoise

- Scientific uncertainty about the superiority of one treatment versus alternative
- Ethical imperative that study participants not be diasadvantaged
- Supports comparative trial design to resolve uncertainty
- If consensus about superiority exists, trial is unethical

From the original "... at the start of the trial, there must be a state of clinical equipoise regarding the merits of regimens to be tested, and the trial must be designed in such a way as to make it reasonable to expect that, if it is successfully conducted, clinical equipoise will be disturbed." (Freedman, 1987)

Stages of trial design in drug development

- Dose-finding trials
  - Phase I
  - Designed to find the best safe dose
  - Traditionally $\leq$ 30 patients
- Safety and efficacy trials
  - Phase II
  - Efficacy signal
  - Traditionally single arm studies with 20 - 80 patients
  - Recent increase in randomized studies for targeted therapies
- Comparative trials
  - Phase III
  - Definitive trial against standard of care
  - Multi-center
  - Hundreds to thousands of patients

Stages of trial design in drug development

- Dose-finding trials
  - Phase I
  - Designed to find the best safe dose
  - Traditionally $\leq$ 30 patients
- Safety and efficacy trials
  - Phase II
  - Efficacy signal
  - Traditionally single arm studies with 20 - 80 patients
  - Recent increase in randomized studies for targeted therapies
- Comparative trials
  - Phase III
  - Definitive trial against standard of care
  - Multi-center
  - Hundreds to thousands of patients

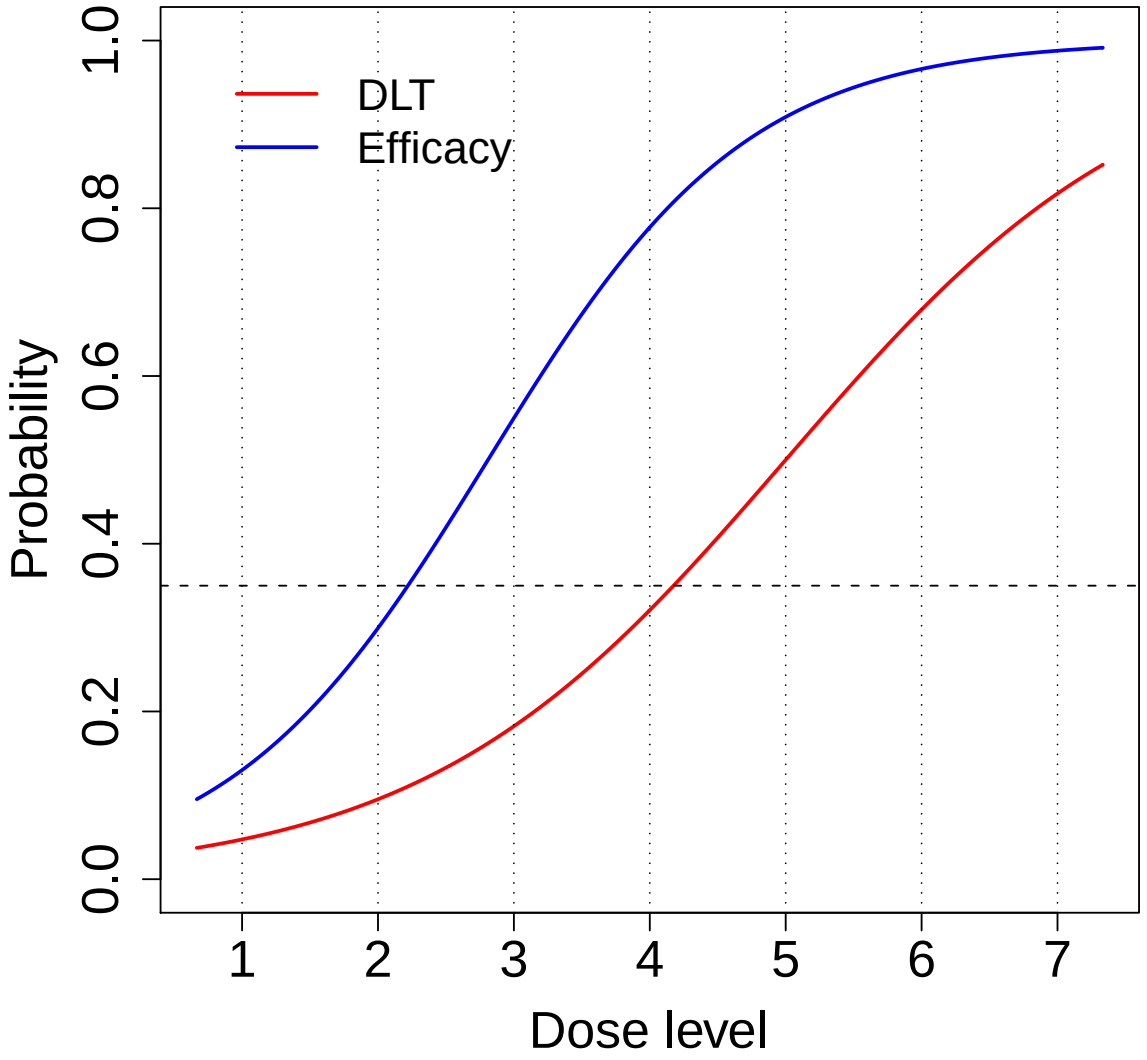
Dose-finding trials

Phase I trial goals

- Classic Phase I trials
  - Cytotoxic agents - “more is better”
  - Find the highest dose that is deemed safe = maximum tolerated dose (MTD)
  - Determine recommended phase II dose (RP2D)
  - Monitor dose-limiting toxicities (DLTs)
  - MTD = highest dose with DLT rate $\leq x\%$ (usually $x\%$ = 20% - 40%)
  - Determine schedule
  - Evaluate safety and toxicity
  - Assess pharmacokinetics
- Newer Phase I trials
  - Molecularly targeted agents
  - Find dose considered safe with optimal biologic/immunologic effect
  - Goal is to optimize ‘biomarker’ response within safety constraints
  - Identify ‘target’ population

Classic phase I assumptions

Efficacy and toxicity both increase with dose

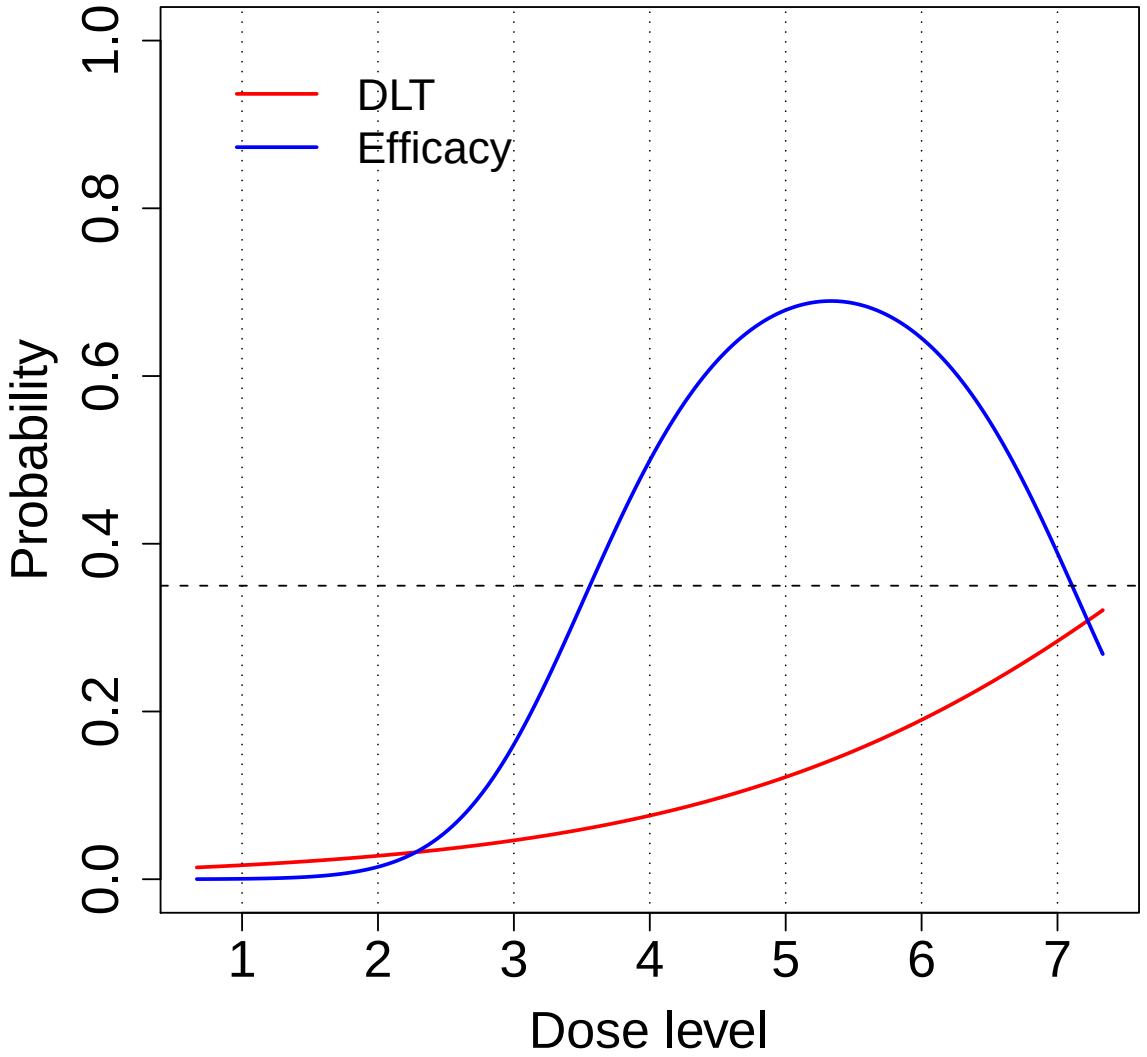


New paradigm for molecularly targeted agents

- Agent selective for a molecular ‘target’
- Disrupt carcinogenesis by interfering with specific pathway
- Can be less toxic than traditional cyotoxic agents
- Implications for phase I design
  - Toxicity may be extremely low
  - Need to ensure agent ‘hits’ the target
  - Efficacy may not increase monotonically with dose
  - MTD being replaced by optimal biologic dose or maximum administered dose

Possible relationships for targeted therapies

-Efficacy not necessarily monotonically increasing with dose
-Low toxicity



,

Phase I designs

- Algorithmic
  - Follows a prescribed set of decision rules for dose-escalation
  - Standard 3+3 design
  - Accelerated titration
- Model-based
  - Uses accumulating toxicity data to determine dose for next patient
  - Continual reassessment method (CRM) (or other flavors - mCRM, TITE-CRM)
- Good overview of Phase I designs and relative merits - Ivy, et al. *Clin. Cancer Res* 2010; 16:1726-36.

Classic phase I "3+3" design

Dose escalation/de-escalation decision rules

For a pre-specified set of doses (usually between 3 and 10), treat 3 patients at dose-level k

1. If 0 of 3 patients experience a DLT, escalate to dose $k+1$
2. If 2 or more of 3 patients experience a DLT, de-escalate to level $k-1$
3. If 1 of 3 patients experiences a DLT, treat 3 more patients at level k
    1. If 1 of 6 experiences a DLT, escalate to dose $k+1$
    2. If 2 or more of 6 experience a DLT, de-escalate to level $k-1$

- MTD = highest dose at which at most 1 out of 6 patients experiences a DLT
- Therefore, target DLT rate is 33%
- Common to include *expansion cohort* at MTD to obtain additional safety data

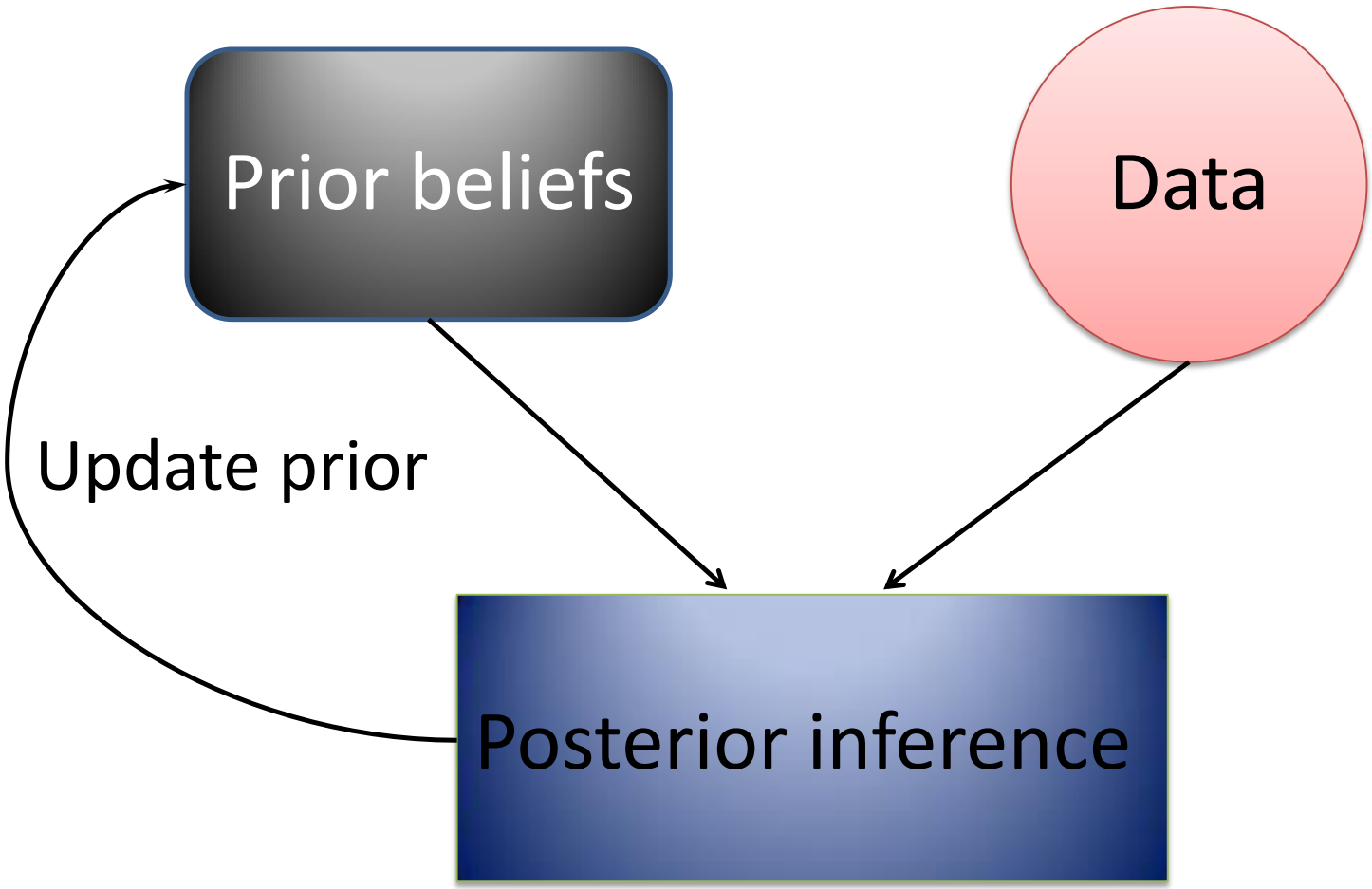
Accelerated titration design

- Original version - Simon, et al. *J Natl Cancer Inst* 1997; 89:1138-47.
- Starts with single patient cohorts
- Approximate doubling of successive doses (96% - 100% dose escalation)
- First DLT or second moderate toxicity in any dose cohort $\Rightarrow$ expand cohort to 3 patients
- Revert to 3+3 design for subsequent cohorts with 40% dose escalation steps
- Advantages over 3+3
  - Requires fewer patients overall (efficiency)
  - Fewer patients treated at non-efficacious doses
  - Increase precision of RP2D

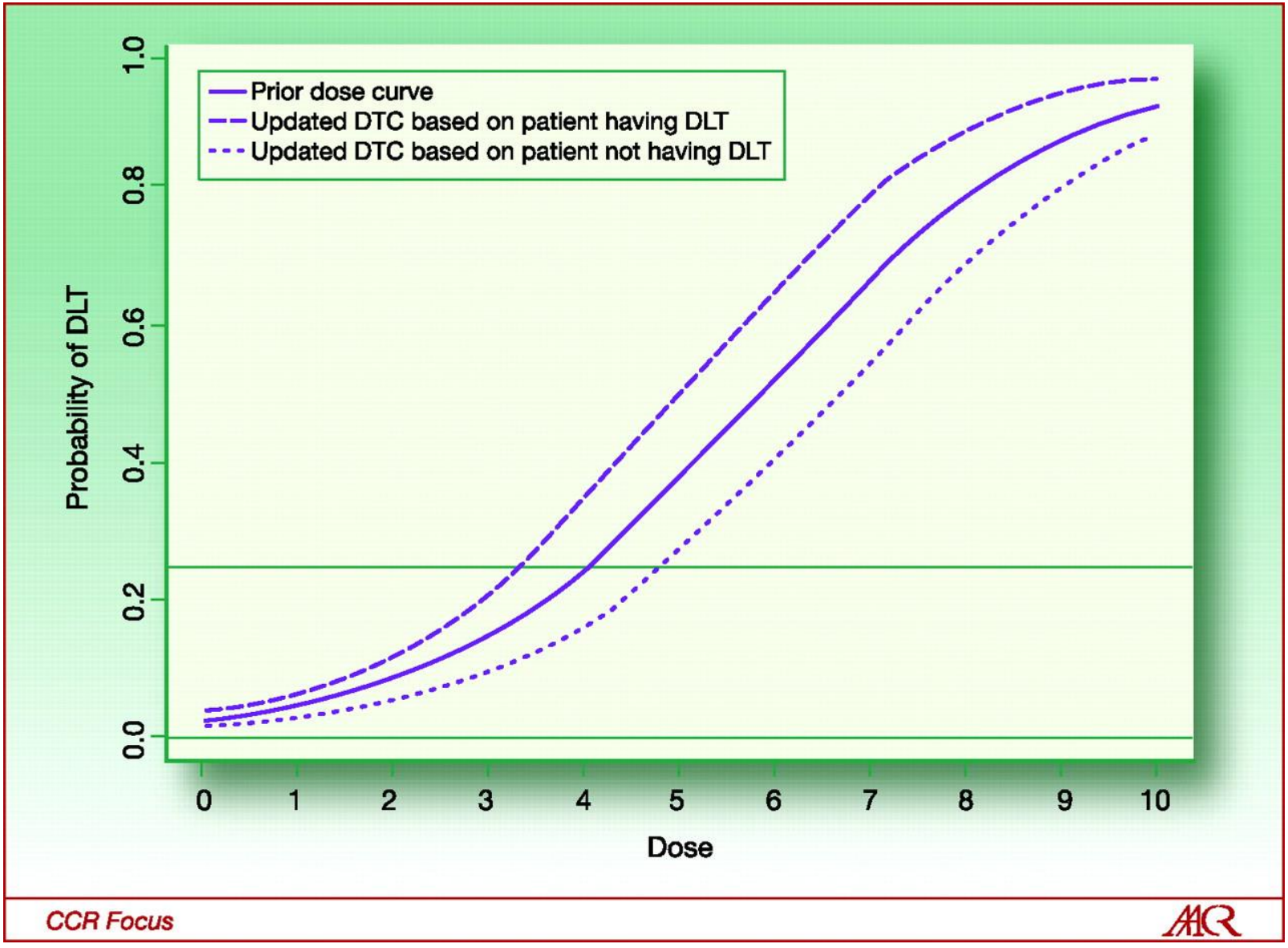
Continual reassessment method

- Original version - O'Quigley, et al. *Biometrics* 1990; 46:33-48.
- Bayesian - uses both accumulating data and prior beliefs to determine actions
- Dose cohorts not specified in advance
- *Learn* best dose assignment for next patient based on toxicity data accumulated at all prior and current dose levels
- Assumes mathematical model relating dose to toxicity
- Requires specification of targeted DLT rate (not 33% by default)
- Patient cohort size must be specified
- Stopping rule - usually based on a pre-specified total sample size

The Bayesian paradigm



CRM dose-toxicity curves



Ivy S P et al. Clin Cancer Res 2010;16:1726-1736

Pros and cons

- Algorithmic designs ...
  - Are simple and easy to understand
  - Are efficient - small sample sizes
  - Can be implemented without involvement of a statistician
  - Enjoy wide-spread acceptance (PRCs, IRBs, FDA)
- Model-based designs ...
  - Are more challenging to understand
  - Treat more patients at efficacious doses
  - Treat fewer patients at toxic doses
  - Identify the MTD with greater precision
  - Require ongoing involvement of statistician
  - Can be difficult to implement at multiple institutions

Safety and efficacy trials

Phase II safety and efficacy trials

- Provide additional information about toxicity profile
- Provide preliminary information on whether a treatment is efficacious
- Small $\Rightarrow$ only large treatment effects are detectable
- Quick $\Rightarrow$ efficacy measured using short-term endpoints
- Phase II endpoints
  - Surrogates for phase III gold standard, overall survival
  - Response is conventional endpoint for cytotoxic agents that cause tumor shrinkage
  - Progression free survival (time from enrollment to documented disease progression) is increasingly the endpoint for MTAs
  - MTAs may not cause tumor shrinkage, but may prolong time to progression

Conventional phase II designs

- Single arm
  - Efficacy rate in study population compared to historical control rate
  - Control rate may not be well-defined
  - Study population may lack comparability to historical control population
- Randomized selection design
  - “Pick the winner”
  - Patients randomized to two (or more) arms
  - No head-to-head comparison of arms, but rather comparisons to null historical rate
  - Goal is to identify best dose/schedule/regimen to take forward into phase III when there is no *a priori* information that one is preferable
- Randomized with control arm
  - Control arm ensures historical rate “on target”
  - Control arm *not* included for head-to-head comparison (due to small sample size)

Conventional phase II design issues

- Issues with respect to molecularly targeted agents
  - Lack of reliable historical control population $\Rightarrow$ no comparator historical rate
  - Response inappropriate endpoint when tumor shrinkage not expected
  - Lack of historical data for appropriate endpoint like PFS
- General issues
  - Lack of randomization $\Rightarrow$ no internal control
  - *Treatment-trial* confounding (Estey and Thall, *Blood* 2003; 102:442-8)
  - Unmeasured *trial effects* (e.g. differences in supportive care, institutional practices, unknown patient characteristics) can not be distinguished from *treatment effects*

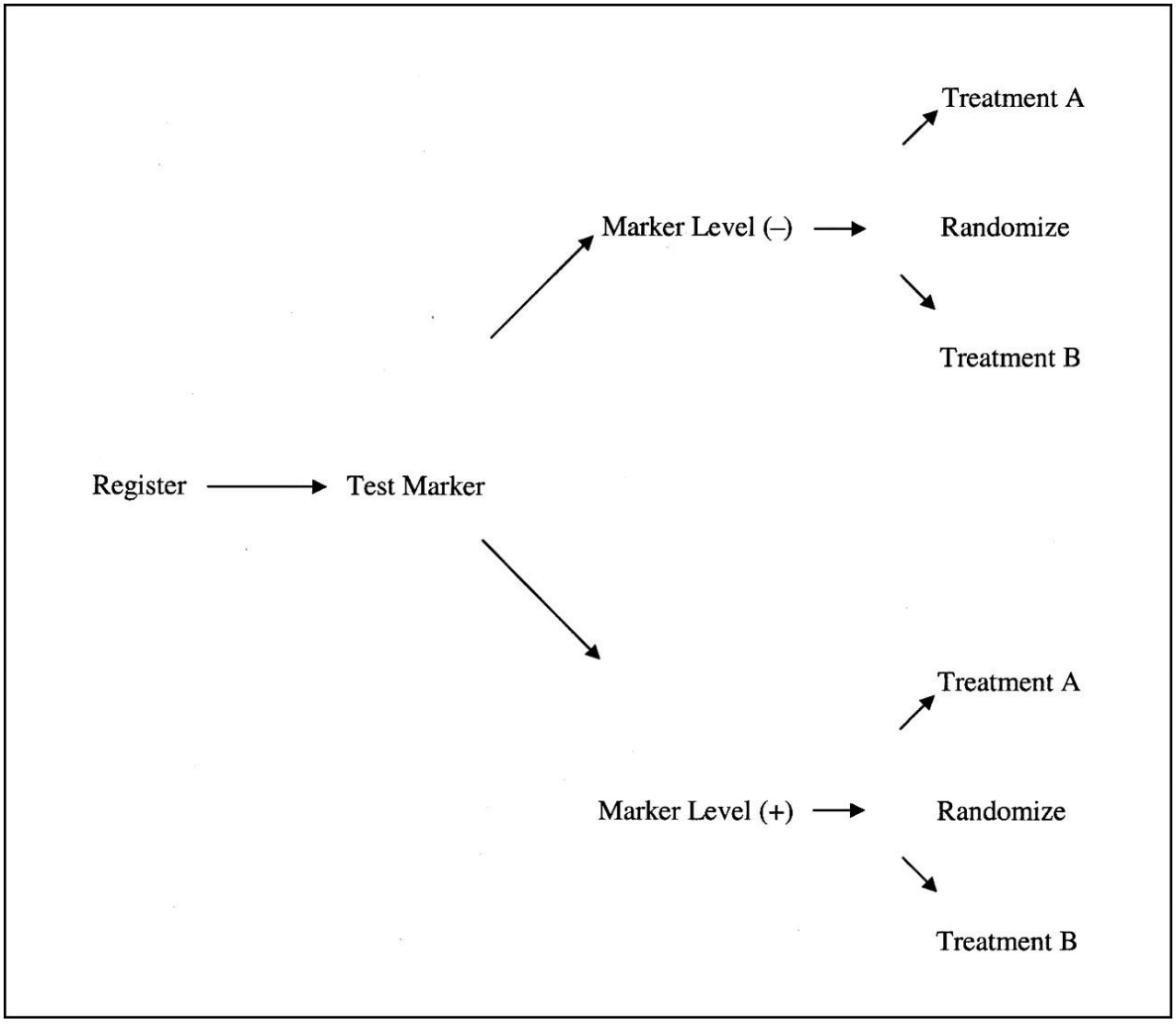
Phase II randomized trials

- Head-to-head comparison
- Increasingly the “norm” for phase II oncology trials
- Requires 2 - 4 times the number of patients as conventional designs
- Increased demand on resources (time and money)
- Standard of care (SOC) + “A” versus SOC (or SOC + placebo)
- Adaptive designs (Gallo et al., *J Biopharm Stat* 2006; 16:275-83)
  - Trial characteristics modified ‘as-you-go’ based on cumulative data
  - Design-based rather than *ad hoc* modifications
  - Trial validity and integrity remains intact
  - Adaptive randomization - increase patients assigned to arm with more efficacious agent
  - Adaptive sample size - re-estimation based on revision of underlying assumptions

Biomarker incorporated designs

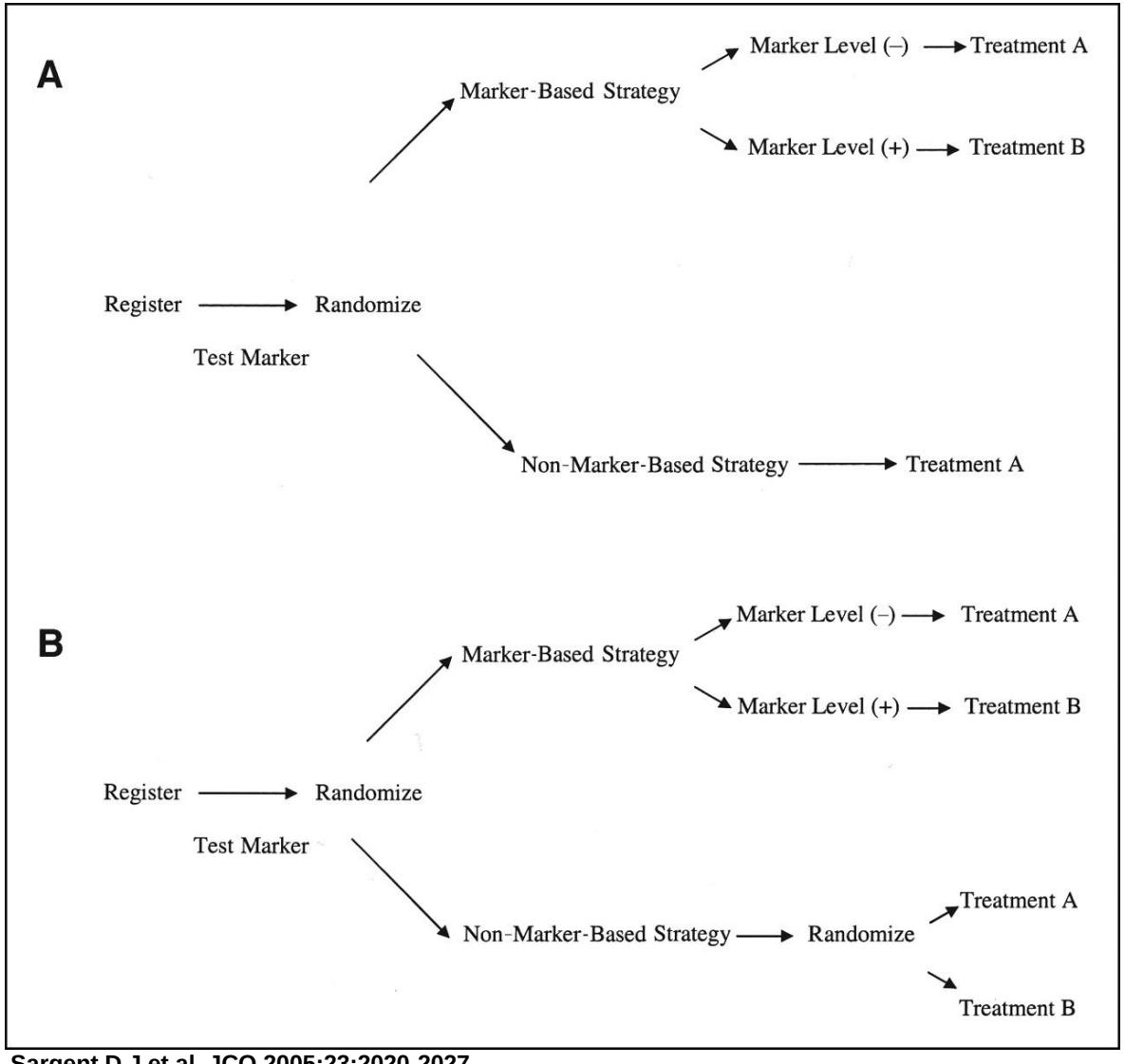
- Randomized phase II designs that utilize biomarker status
- Focus on predictive markers
- Often goal is to assess treatment efficacy and validate marker
- Sargent et al, *JCO* 2005; 23:2020-27
- Selected designs
  - Enroll patients most likely to benefit - marker positive patients
  - Requires biologic accuracy of selected marker
  - Reliable marker measurement
  - Fails to identify other patient populations who may benefit
  - Does not allow assessment of marker-outcome association
- Unselected designs
  - Marker-by-treatment interaction design
  - Marker-based strategy design

Marker-by-treatment interaction design



Sargent D J et al. JCO 2005;23:2020-2027

Marker-based strategy design



Sargent D J et al. JCO 2005;23:2020-2027

Statistical and design considerations

Hypothesis testing

- Didactic example - Randomized phase II trial of paclitaxel/carboplatin $\pm$ enzastaurin (PCE/PC) in advanced ovarian cancer patients (Vergote et al., *JCO* 2013; 31:3127-3132)
- To compare competing treatments, we begin with rival hypotheses of their relative effect.
  - *Null hypothesis* - statement of no effect (H_0)
  - *Alternative hypothesis* - statement reflecting conclusion if trial results suggest null is implausible (H_A)

H_0: Advanced ovarian cancer patients derive no additional clinical benefit from the addition of enzastaurin to paclitaxel/carboplatin relative to those treated with paclitaxel/carboplatin alone.

H_A: Advanced ovarian cancer patients derive significant clinical benefit from the addition of enzastaurin to paclitaxel/carboplatin relative to those treated with paclitaxel/carboplatin alone.

Formalizing hypothesis testing

- ‘Clinical benefit’ needs to be articulated in terms of a measureable quantity
- Depends on the trial’s primary endpoint
- If endpoint is response, benefit measured by difference in response rate
- If endpoint is a time-to-event measure, benefit measured by *hazard ratio* (HR)

Hazard ratio detour

- HR = $\frac{\text{risk of event in trt grp}}{\text{risk of event in ctl group}}$
- HR $>$ 1 indicates increased risk of event in treatment group relative to control
- HR $<$ 1 indicates decreased risk of event in treatment group relative to control
- HR = 1 indicates equivalent risk of event in treatment and control groups
- Interpretation
  - HR = 0.80 $\Rightarrow$ "*There is a 20% reduction in the risk of death (progression, recurrence) comparing treated to control patients.*"
  - HR = 1.20 $\Rightarrow$ "*There is a 20% increase in the risk of death (progression, recurrence) comparing treated to control patients.*"

Back to our hypotheses

If endpoint is response ...

$$H_0: p_{\text{PCE}} = p_{\text{PC}}$$

$$H_A: p_{\text{PCE}} > p_{\text{PC}}$$

where p_{PCE} is the response rate in the PCE arm and p_{PC} is the response rate in the PC arm.

If endpoint is a time-to-event measure

$$H_0: HR_{\text{PCE:PC}} = 1$$

$$H_A: HR_{\text{PCE:PC}} < 1$$

Conducting inference and quantifying evidence

- Based on the evidence in the data, one of two decisions is made
  1. Reject H_0 in favor of $H_A \Rightarrow$ sufficient evidence to rule out the null
  2. Fail to reject $H_0 \Rightarrow$ insufficient evidence to rule out the null
- Evidence is quantified using a *P-value*
- If the null hypothesis is true, it is possible to observe what looks like a treatment effect by pure chance
- P-value is the probability of observing a result at least as extreme as what was observed in the current study if the null is true (i.e. if there really is no difference)
- P-value quantifies the probability of a spurious finding
- Usually require $p\text{-}value < 0.05$ to reject H_0 and declare *statistical significance*
- $p\text{-}value \geq 0.05 \Rightarrow$ fail to reject H_0 and conclude we can't rule out chance as a plausible explanation for any observed difference

Confidence intervals

- Provides range of values quantifying uncertainty associated with an estimated parameter
- The wider the interval the greater the uncertainty in the estimate
- The more narrow the interval the greater our faith in the estimate
- Typically report 95% CI
- The “95%”-part of the interval means that if you were able to replicate the exact same study an infinite number of times, 95% of the resulting CIs would contain the true parameter of interest (not a particularly practical interpretation)

Putting some of this together ...

In the randomized phase II PCE vs PC trial in ovarian cancer,

"A comparison of ... PFS for the two regimens ... was not statistically significant ($P = 0.367$). HR at the time of final analysis was 0.80 (95% CI, 0.50 to 1.29)."

Interpretation and conclusions

- Enzastaurin does not significantly increase time to progression when added to a regimen of paclitaxel/carboplatin in advanced ovarian cancer patients ($P = 0.367 \geq 0.05$)
- Adding enzastaurin to paclitaxel/carboplatin therapy conferred a 20% reduction in the risk of progression (HR = 0.80), although this finding was not significant. That is to say, although we observed an improvement, we can not rule out chance as a plausible explanation for the reduction in risk.
- Our best estimate of the hazard ratio is 0.80, but the data are also consistent with a HR ranging from 0.50 to 1.29 (95% CI, 0.50 to 1.29). Thus, the data are consistent with equivocal conclusions - a hazard ratio of 0.50 indicates a 50% reduction in the risk of progression while a hazard ratio of 1.29 indicates a 29% increase in the risk of progression.

Actions and errors

Of course it is always possible to make a mistake. The table below details the two types of errors one can commit in conducting a statistical test.

	H_0 true	H_A true
Reject H_0	Type I error	No error
Fail to reject H_0	No error	Type II error

- A type I error is the probability of rejecting the null given that the null is true.
- We denote the probability of a type I error as α.
- A type II error is the probability of failing to reject the null given that the alternative is true.
- We denote the probability of a type II error as β.

Power

	H_0 true	H_A true
Reject H_0	α	1 - β = Power
Fail to reject H_0	1 - α	β

The power of a test is the probability of correctly rejecting the null in favor of the alternative. A well-designed study will strike a balance between acceptable levels of type I error (usually 0.05) and power (often set at a minimum of 0.8).

Remember ...

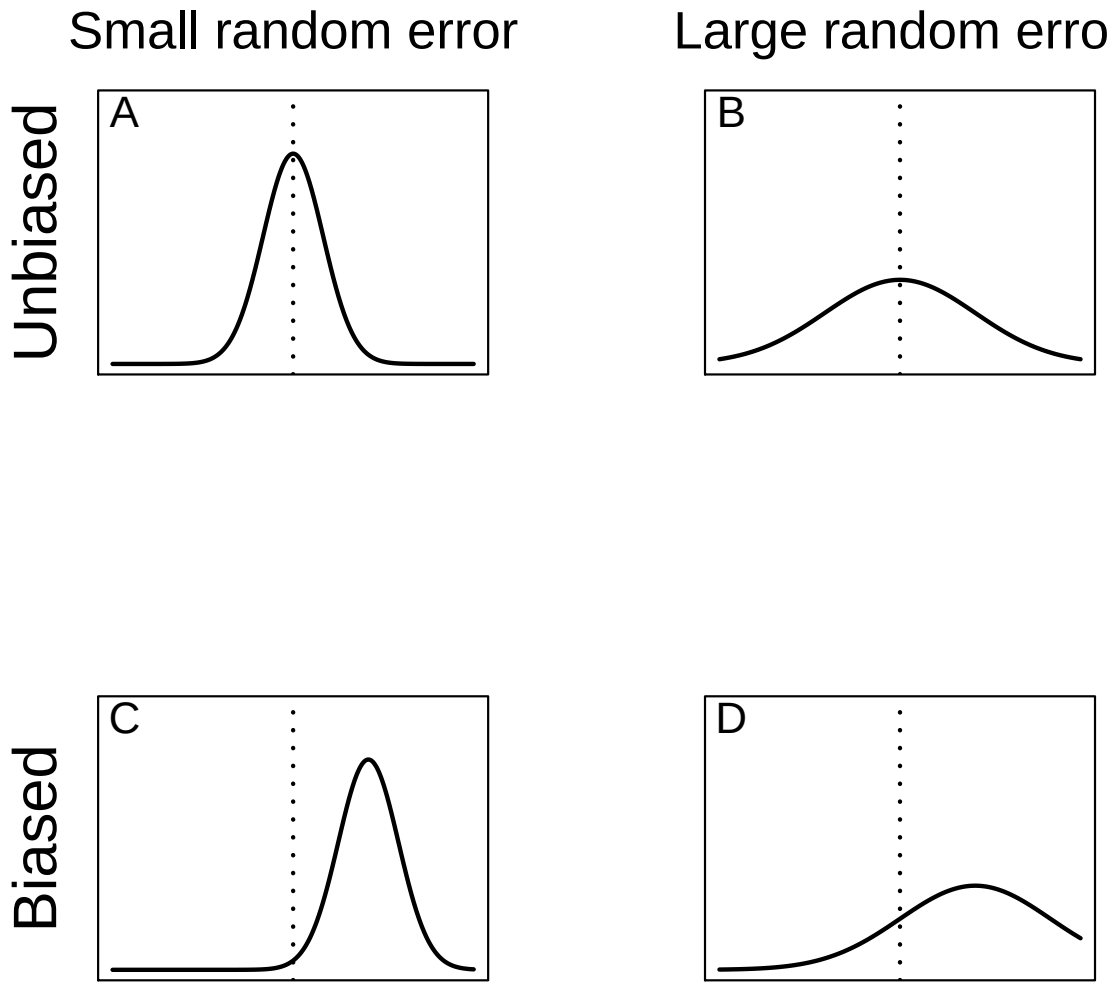
- A type I error is rejecting the null when you shouldn't
- Power is the probability of rejecting the null when you should

Planned early data looks

- Early stopping for efficacy
  - Is there sufficient evidence to conclude treatment is significantly better?
  - Ethical imperative not to continue to randomize patients to a therapy known to be inferior
  - Inflates type I error rate
- Early stopping for futility
  - No hope of being able to demonstrate treatment efficacy
  - Sufficient data to answer scientific question
  - Unethical to continue to randomize patients in a trial with no additional benefit
  - Inflates type II error rate (decreases power)
- Early stopping for safety

Bias and error

Control of *bias* and reduction of *random error* are two major objectives in statistical design considerations



Sources of bias

1. *Selection bias*
   - Patients in arms systematically different with respect to prognostic factors
     - Bias the observed treatment effect
     - Can influence internal validity
   - Study cohort not representative
     - Can influence external validity
     - Compromises generalizability
2. *Treatment/procedure selection bias*
   - Healthier patients selected for a particular treatment
   - Systematic difference in composition of treatment groups
   - Can bias treatment difference

Sources of bias (cont.)

3. Postentry exclusion bias

- Inappropriate exclusion of eligible and enrolled subjects from the analysis
- Exclusion often due to seemingly reasonable clinical reasons
- Breaks the ‘experimental paradigm’
- Example - subjects that fail to complete therapy are excluded from the analysis
- Example - subjects that die due to ‘other’ causes are excluded from an analysis of overall survival

Sources of bias (cont.)

4. Selective loss of data

- Loss of data resulting from unworkable or suboptimal outcomes or errors in study conduct
- Endpoint poorly selected for patient population
- Frequency of follow-up inappropriate for assessment of desired endpoint or is not followed as specified in protocol
- Example - patient population is seriously ill cohort and endpoint is based on patient self-report; endpoint may suffer from survivor bias
- Example - endpoint is time to progression; follow-up with patients every 6 months may be inadequate for accurate assessment

Sources of bias (cont.)

5. *Assessment bias*
   - Patient self-assessment lacks objectivity
   - Clinician assessment can be influenced by expectation of treatment effect
   - Can bias endpoint in direction of prior expectation
6. *Uncontrolled confounders*
   - Confounder is a variable that masks the true treatment effect
   - Common confounders are age, race, gender, disease severity, comorbidities
   - Example - treatment arm is significantly younger than placebo arm, and outcomes in older patients are more severe

Controlling for bias by design - Randomization

- Patients randomly assigned to treatment
- Controls for:
  - Selection bias
  - Treatment bias
  - Uncontrolled confounders
- Types of randomization
  - Simple
  - Permuted block
  - Stratified permuted block
  - Adaptive
  - Group
- Allocation ratio (1:1, 2:1, etc.)

Blinding

- *Blinding* = treatment masking
- Treatment masked from the patient - single blind
- Treatment masked from both the patient and the study personnel - double blind
- Blinding controls for
  - Treatment/procedure bias
  - Assessment bias
- No blinding for members of DSMB
- Blinding isn't always possible
  - Drugs being compared have different modes of delivery - infusion versus tablet
  - Blinding in trials of devices or surgical procedures is difficult or impossible

Study populations

Intention-to-treat (ITT) is the idea that patients in a randomized clinical trial should be analyzed as part of the treatment group to which they were assigned, even if they did not actually receive the intended treatment. For assessing efficacy, analysis of the ITT population is preferred.

Treatment received (TR) is the idea that patients should be analyzed according to the treatment actually given, even if the randomization calls for something else. For assessing safety, analysis of the TR population is preferred. This is sensible since we want to attribute any severe adverse events to the treatment actually received.

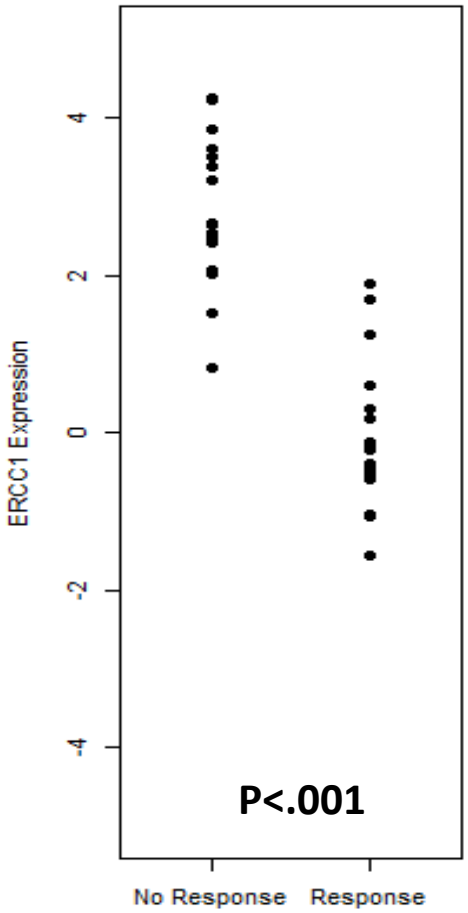
Efficacy based on the ITT population

- Many factors contribute to a patient's inability to complete the intended therapy or patient adherence
  - Side effects
  - Disease progression
  - Patient/physician preference for a different treatment
- Failure to complete therapy as randomized is almost always an outcome of the study itself
- If ITT population is not analyzed, can result in post-entry selection bias
- Results from efficacy analysis using ITT population is a test of treatment policy/program effectiveness

Correlative biomarker studies

- Secondary or exploratory aims of trial
- Examine association of predictive marker with clinical endpoint
- Example (used with permission) from M Regan, DFCI
- Phase II trial to evaluate response of patients with muscle-invasive urothelial cancer treated with neoadjuvant dose-dense methotrexate, vinblastine, doxorubicin, cisplatin (ddMVAC), followed by radical surgery with curative intent (Choueiri, ASCO 2013, 4530)
- Single-arm, 2-stage design (futility look) - $\alpha = 0.1$, power = 0.85, $n = 37$
- Correlative objective: Investigate tumor expression levels of DNA repair genes (e.g. ERCC1) in relation to response
- High tumor tissue levels of ERCC1 mRNA have been associated with clinical resistance to cisplatin-based chemotherapy in ovarian, gastric, cervical, colon and NSCLC patients

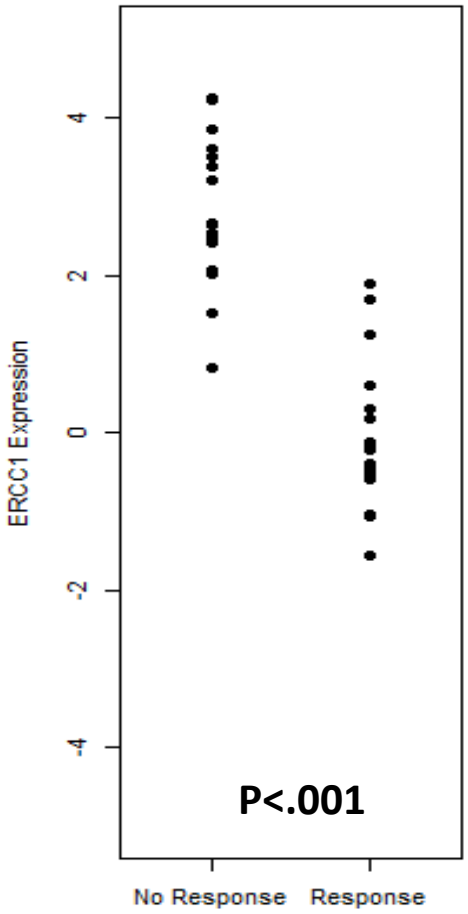
ddMVAC example: ideal versus reality



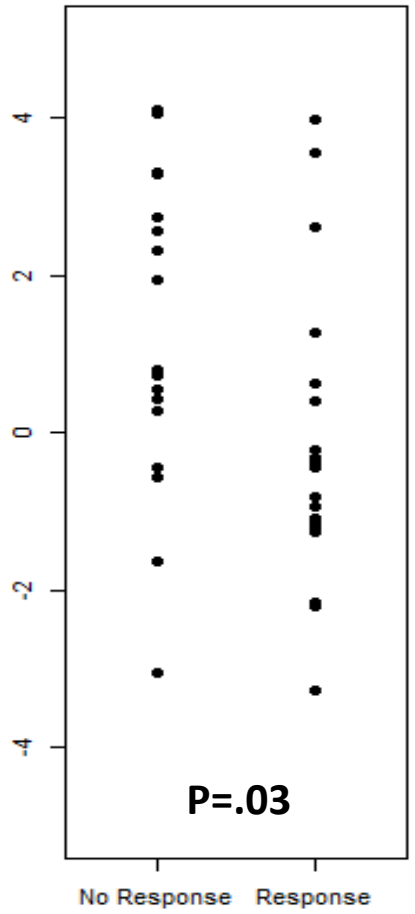
***Ideal* (n=37)**
55% response

Result the investigator anticipates: large, clear difference

ddMVAC example: ideal versus reality



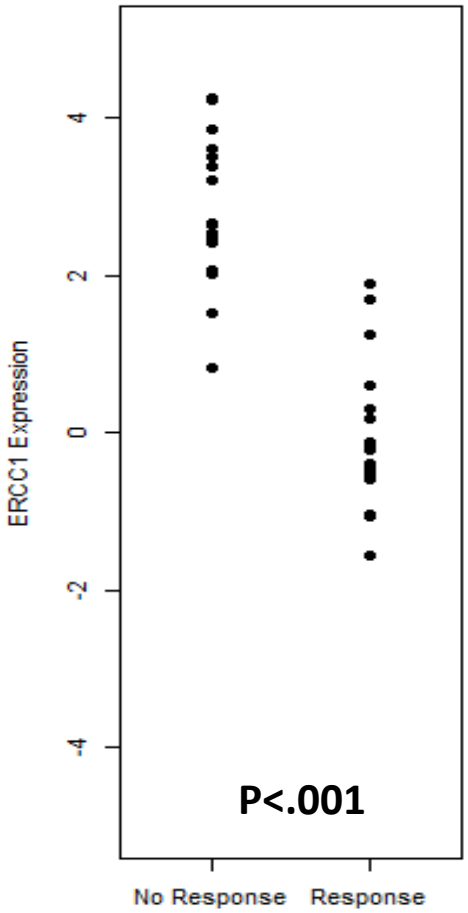
***Ideal* (n=37)**
55% response



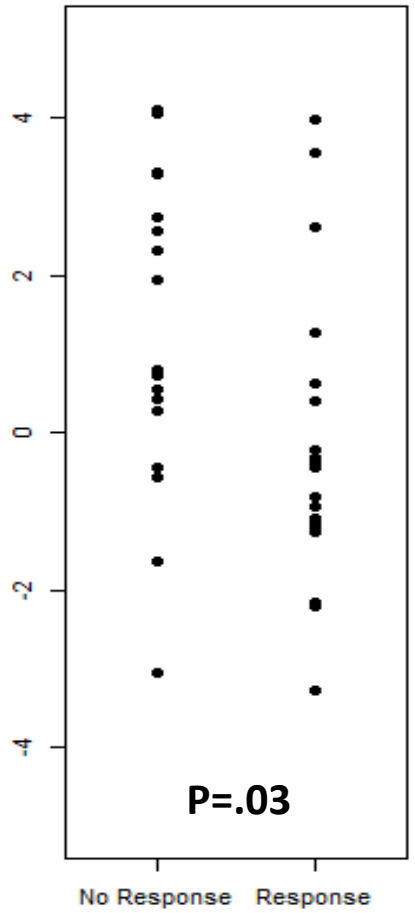
Less ideal (n=37)
55% response

Statistician's anticipation: smaller difference, greater variability

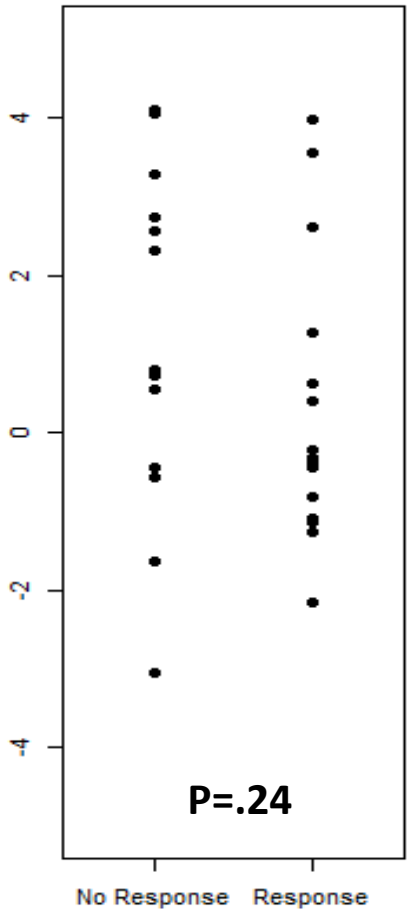
ddMVAC example: ideal versus reality



***Ideal* (n=37)**
55% response



Reality (n=37)
55% response



Reality (n=27) successful assays
55% response