2024 Summer Institute In Statistics for Clinical & Epidemiological Research

Module 3:

Design, Conduct, and Analysis of Randomized Clinical Trials with Time to Event Primary Endpoints

Lecture 17:

Randomization

Scott S. Emerson, M.D., Ph.D.
Professor Emeritus of Biostatistics
University of Washington

1

Comment re Single Arm Trials

"There are only two types of researchers:

- those with a lot of enthusiasm and no controls, and
- those with a lot of controls and no enthusiasm."

(unknown)

2

2

Real-life Examples

- Effects of arrhythmias post MI on survival
  - Observational studies: high risk for death
  - CAST: anti-arrhythmics have higher mortality
- Effects of beta-carotene on lung CA and survival
  - Observational studies: high dietary beta carotene has lower cancer incidence and longer survival
  - CARET: beta carotene supplementation in smokers leads to higher lung CA incidence and lower survival
- Effects of hormone therapy on cardiac events
  - Observational studies: HT has lower cardiac morbidity and mortality
  - WHI: HT in post menopausal women leads to higher cardiac mortality

3

3

No Comparison Group

- Appropriate when an absolute criterion for treatment effect exists
- Single arm clinical trial
  - Cohort design
  - Includes “pre-post” designs
- (Rarely do such absolute criteria exist. Instead, we are really invoking the use of results from previous investigations.)

4

4

Historical Controls

- Single arm clinical trial
- Compare results to
  - Absolute criterion derived from historical trials
    - Dishonest: Use only one-fourth the sample size
  - Sample from historical clinical trial (better)
    - More honest: Maybe only save half the sample size
- However, the validity of such methods is heavily dependent upon the historical trial being comparable in every way
  - No changes in comparison treatment
  - No changes in definition of study population
  - No changes in ancillary treatments
  - No changes in measurement of treatment outcome

5

5

Attempted Statistical Solutions

- Adjustment for confounders or propensity score analyses suffer from drawbacks noted by Byar (Biometrics, 1980) and Simon (Ca Treat Rep, 1982):
  - The variables that are measured and properly recorded typically explain only a small percentage in the variability in treatment group membership and treatment outcome.
    - That is, the regression models used have a very low R^2, thus our ability to have properly matched groups is rather low.

6

6

Example: Propensity Scores with WHI Data

- Propensity score analyses to reconcile RCT, observational data?

Open Access Research

BMJ open

Comparing hormone therapy effects in two RCTs and two large observational studies that used similar methods for comprehensive data collection and outcome assessment

Arthur Hartz,[1] Tao He,[2] Robert Wallace,[3] John Powers[4]

To cite: Hartz A, He T, Wallace R, *et al.* Comparing hormone therapy effects in two RCTs and two large observational studies that used similar methods for comprehensive data collection and outcome assessment. *BMJ Open* 2013;**3**:e002556. doi:10.1136/bmjopen-2013-002556

► Prepublication history and additional material for this paper is available online. To view these files please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2013-002556).

Received 8 January 2013

ABSTRACT

Objectives: Prospective observational studies (OSs) that collect adequate information about confounders can validly assess treatment consequences. However, what constitutes adequate information is unknown. This study investigated whether the extensive information collected by the Women's Health Initiative (WHI) in two OSs and two randomised controlled trials (RCTs) was adequate.

Design: Secondary analysis of WHI data. Cox regression was used to select from all baseline risk factors those that best predicted outcome. Cox regression that included these risk factors was used for two types of analyses: (1) comparing RCT and OS assessments of the effects of hormone therapy on outcome for participants with specific characteristics and (2) evaluating whether adjustment for measured confounders could eliminate outcome differences among datasets.

Setting: The WHI included more than 800 baseline risk factors and outcomes during a median follow-up of 8 years.

ARTICLE SUMMARY

Article focus

- Observational studies (OSs) are frequently used to compare outcomes of patients who choose different treatments.
- Results of OSs may be invalid because of confounding due to an association between patient risk and treatment choice.
- The present study assessed whether the extensive information collected by the Women's Health Initiative (WHI) was adequate to eliminate confounding and give valid results.

Key messages

- The effects of hormone therapy on stroke and myocardial infarction differ for OSs and randomised controlled trials even after taking advantage of extensive participant information to remove confounding and to select similar participants.

7

7

Example: Propensity Scores with WHI Data

- Propensity score analyses to reconcile RCT, observational data?

Hartz A, He T, Wallace R, *et al. BMJ Open* 2013;**3**:e002556. doi:10.1136/bmjopen-2013-002556

Comparing hormone therapy effects in two RCTs and two large observational studies

Table 2 Risk-adjusted HRs for hormone therapy in different datasets

		Myocardial infarction		Stroke	
Dataset	**HT type**	**HR**	**95% CI**	**HR**	**95% CI**
WHI OS	Any E	0.83	(0.72 to 0.95)	0.85	(0.70 to 1.03)
	E+P	0.86*	(0.70 to 1.05)	0.82*	(0.65 to 1.04)
	E-alone	0.80†	(0.69 to 0.94)	0.88‡	(0.71 to 1.11)
Diet RCT	Any E	0.75	(0.62 to 0.89)	1.04	(0.80 to 1.37)
	E+P	0.96	(0.75 to 1.22)	1.00	(0.72 to 1.39)
	E-alone	0.65†§	(0.53 to 0.81)	1.07	(0.79 to 1.45)
HT RCT	Any E	1.18	(0.99 to 1.41)	1.29	(1.05 to 1.58)
	E+P	1.30	(1.02 to 1.65)	1.34	(1.02 to 1.77)
	E-alone	1.05	(0.81 to 1.36)	1.23	(0.91 to 1.67)

*Differs from the comparable RCT HR at the p<0.01 level.
†Differs from the comparable RCT HR at the p=0.02 level.
‡Differs from the comparable RCT HR at the p=0.06 level.
§Differs from 1.00 at the p<0.0001 level.
Any E, E+P or E-alone; E-alone, oestrogen alone; E+P, oestrogen plus progesterone; HT, hormone therapy; OS, observational study; RCT, randomised controlled trial; WHI, Women's Health Initiative.

8

Internal Controls

- Each subject serves as his/her own control
  - Different treatments at different times
    - washout period necessary
  - Different treatments for different parts of body
    - eye diseases, skin diseases
    - need to avoid cross-contamination
- In a “cross-over design”, order of treatments should be randomized
  - Contrast with “before-after” single arm trial

9

9

Concurrent control group

- Two or more treatment arms
  - Randomized
  - Placebo or standard therapy
  - Active treatments
    - Sometimes consider equivalence
  - Multiple levels of same treatment
    - Stronger evidence sometimes obtained from dose-response
      - Koch’s postulates
    - Identifying optimal dose

10

10

Cause and Effect

- Necessary conditions for establishing cause and effect of a treatment
  - The treatment should precede the effect
    - Beware protopathic signs
      - Marijuana and risk of MI within 3 hours
  - When comparing groups differing in their treatment, the groups should be comparable in every other way (at baseline)

11

11

Major Scientific Tool

- Randomization is the major way in which cause and effect is established
  - Ensures comparability of populations
    - Each treatment group drawn from same population
    - Differences in other prognostic factors will only differ by random sampling
      - Provides balance on the total effect of all other prognostic factors
      - May not provide balance on each individual factor
- NB: Sequential allocation of patients is not randomization
  - Possible time trends in recruitment, treatments, etc.

12

12

Blinding

- In studies with concurrent comparison groups, blinding of treatment assignment can minimize bias
  - Single blind experiments:
    - Participant is unaware of treatment assignment
  - Double blind experiments:
    - Neither participant nor provider know treatment assignment
  - Triple blind experiments:
    - Monitoring committee also semi-blinded
  - Blinded evaluation of outcomes
    - Outcomes for each patient defined by blinded evaluator

13

13

Advantages

- Blinding can serve to
  - Minimize "placebo effect": A participant being treated does better than one not being treated, irrespective of the actual treatment
  - Minimize investigator bias in assessing
    - accrual to study
    - adverse events
    - treatment outcomes

14

14

Impact on Treatment Definition

- In human experimentation, we never test a treatment
  - We may not ethically force people to continue a therapy
  - It may not be medically advisable to even want a patient to continue
    - Patients may discontinue a therapy due to headache
    - If forced to continue, those patients may have CVA
- Instead we test a treatment strategy
  - We prescribe an initial treatment
  - Patients may also receive ancillary treatments
    - These may be precipitated by experimental therapy
  - Patients may progress to other therapies

15

15

Definition of Treatments

- Full description
  - Formulation of treatment
  - Dose, administration, frequency, duration
    - Rules for responsive dosing (e.g., insulin)
    - Include plans for
      - Treatment of adverse events
      - Dose reduction
      - Dose discontinuation
  - Ancillary treatments
    - Prescribed vs allowed vs prohibited
      - (Distinguish safety issues from efficacy issues)

16

16

Special Issues

- Ultimately, the scientific credibility of the clinical trial stems from our ability to assign a treatment to the participants
- Ideally we do this in a random fashion
  - Sequential allocation is not random
- At a given point in time, we can only assign a strategy
  - Competing risks may make treatment impossible
  - Intervening events may change indications
  - Informed consent can be withdrawn
- We must avoid ruining the comparisons of strategies
  - Naïve attempts to compare "treatment" may ruin our ability to assess what really can be tested

17

17

Ramifications

- Possible actions on progression
- Stay the course
  - "Progression" dichotomizes a continuous process
  - Treatment may be delaying that process
- Advance to other therapies
  - Ideally the same for both treatment arms
- Cross-over to other arm
  - Sometimes motivated to increase sample treated
  - A huge scientific mistake but
    - Ethics sometimes demands it
      - PA catheterization vs central line
      - Pemetrexed vs docetaxel

18

18

Comments

- Can there be a noncompliant subject?
- Experimentally: NO
  - By definition, all patients are following our strategy of having been told what treatment to take
    - Clearly addresses effectiveness questions
    - If efficacy had been our goal:
      - Exclude noncompliant patients as much as possible
      - Increase sample size to deal with attenuation
- Safety: MAYBE
  - We do have to worry that adherence to treatment strategy may change after reporting efficacy
    - We will only have tested safety under the compliance actually achieved
  - Measuring compliance is important for interpretation

19

19

Ramifications

- An important distinction needs to be made between
  - "Stopping study drug"
    - This may happen due to
      - Adverse events
      - Progression
      - Study burden
    - While we hope for high compliance
      - Badgering patients to remain on therapy can lead to worse adverse events or the quitting the study
    - In the event of stopping study drug, all follow-up of primary outcomes should proceed as planned
  - "Withdrawing consent"
    - No further data will be available

20

20

Randomization Strategies

- Complete randomization (CRD)
- Blocked randomization
  - Ensure balance after every *k* patients
  - Ensure closer adherence to randomization ratio
  - Undisclosed block sizes to prevent bias
- Stratified randomization
  - Separately within strata defined by strong risk factors
    - Lessens chance of randomization imbalance
  - Need to consider how many variables can be used
- Dynamic randomization
  - Adaptive randomization to achieve best balance on marginal distribution of covariates

21

21

Why Randomize?

Common Statistical Analysis Models

Where am I going?

- The scientific question posed by a clinical trial is typically translated into a statistical comparison of probability distributions
  - Unadjusted or adjusted comparison of summary measures
- We will need to describe the statistical implications of any randomization strategy in the context of statistical analysis model
  - Notation for regression on means, odds, or hazards

22

22

Summary Measures

- The measures commonly used to summarize and compare distributions vary according to the types of data
  - Means: binary; quantitative
  - Medians: ordered; quantitative; censored
  - Proportions: binary; nominal
  - Odds: binary; nominal
  - Hazards: censored
    - hazard = instantaneous rate of failure

23

23

Analytic Randomization Models

- Population model
  - Ensures treatment arms drawn from same population initially
  - Test weak null hypothesis of no treatment effect on summary measure of interest
    - E.g., test of equal mean outcome
    - Can allow for treatment differences between arms on other aspects of outcome distribution
- Randomization model
  - Conditions on the sample obtained
    - E.g., permutation tests
    - Pretends that all outcomes were pre-ordained absent a treatment effect
  - Tests strong null hypothesis of no treatment effect whatsoever
    - Under the null hypothesis, any difference in outcome must have been randomization imbalance

24

24

Comments: Strong vs Weak Null

- Logical implications
  - Strong Null ➔ Weak Null
  - Rejection of Weak Null ➔ Rejection of Strong Null
- Advantages / Disadvantages of Strong Null
  - Can always test strong null via permutation tests in fixed sample
    - Sequential sampling poses problems
  - Assumption of strong null not in keeping with scientific method
    - Assumptions are more detailed than primary question
      - Primary question usually about first moment
      - Semiparametric assumptions are about all moments
    - Consider bone marrow transplantation

25

25

Comments: Choice of Analytic Models

- First choice: Population model
  - Randomization model does not typically allow testing of nonzero null hypotheses (e.g, noninferiority)
  - Randomization model does not allow distribution-free estimation of confidence intervals
    - For CI, we must know distribution under alternatives
- But the randomization model is an important fall back position
  - I generally feel uncomfortable in settings where a population model rejected a weak null but a randomization model could never reject the strong null
  - (cf: Deterministic minimization methods)

26

26

Points Meriting Repeated Emphasis

- Randomization is our friend…
  - If we randomize, we do not (on average) need to worry about differences between the treatment groups with respect to factors present at time of randomization
    - Any difference in outcomes can be attributed to treatment
      - Again, recognize that treatment can lead to differential use of other ancillary treatments, however
- But like all friends, we must treat it with respect.
  - We must analyze our data in groups defined at the time of randomization
    - Discarding or missing data on randomized subjects may lead to bias
      - It certainly leads to diminished scientific credibility

27

27

Impact on Data Analysis

- In presence of randomized treatment assignment
  - Intent to treat analysis (ITT)
  - Based on randomization
    - "Modified ITT" acceptable for efficacy?
      - Efficacy within strata identified pre-randomization
      - Safety in all subjects
  - Science: Population model (not randomization model)
    - My view: "Permutation Tests Considered Harmful"

28

28

Points for Further Elucidation

- Confounding not an issue (on average)
  - P value measures probability of observed effects occurring due only to randomization imbalance
- Gain precision if
  - Control important prognostic variables, or
  - Adjust for stratification variables
- Subgroup analyses
  - If effect modification is concern

29

29

Regression Models

- According to the parameter compared across groups
  - Means → Linear regression
  - Geom Means → Linear regression on logs
  - Odds → Logistic regression
  - Rates → Poisson regression
  - Hazards → Proportional Hazards regr
  - Quantiles → Parametric survival regr

30

30

Statistical Models

- How are (semi)parametric assumptions really used in statistical models?
  - Choice of functional for comparisons
    - (Should use scientific loss function)
  - Formula for computing the estimate of the functional
    - (Should be distribution-free)
  - Distributional family for the estimate
    - (CLT: Typically asymptotically normal – like it or not)
  - Mean-variance relationship across alternatives
    - (This is what matters)
  - Shape of distribution for data
    - (Only matters for prediction)

31

31

"Everything is Regression"

- The most commonly used two sample distribution-free tests are special cases of regression

- Regression with a binary predictor
  - Linear → t test
  - Logistic → chi square (score test)
  - Proportional hazards → logrank (score test)

32

32

General Regression

- General notation for variables and parameter

Y_i	Response measured on the ith subject
X_i	Value of the POI for the ith subject
$W_{1i}, W_{2i}, \ldots$	Value of adjustment variables for the ith subject
θ_i	Parameter of distribution of Y_i

  - The parameter might be the mean, geometric mean, odds, rate, instantaneous risk of an event (hazard), etc.

33

33

Interpretation of Slopes

- Difference in interpretation of slopes

$$\text{Unadj Model}: \quad g\left[\theta \,|\, X_i\right] = \beta_0 + \beta_1 \times X_i$$

  - β_1 = Compares θ for groups differing by 1 unit in X
    - (The distribution of W might differ across groups being compared)

$$\text{Adj Model}: \quad g\left[\theta \,|\, X_i, W_i\right] = \gamma_0 + \gamma_1 \times X_i + \gamma_2 \times W_i$$

  - γ_1 = Compares θ for groups differing by 1 unit in X, but agreeing in their values of W

34

34

Defining "Contrasts"

- Define a comparison across groups to use when answering scientific question
- If straight line relationship in parameter, slope for POI is difference in parameter between groups differing by 1 unit in X when all other covariates in model are equal
- If nonlinear relationship in parameter, slope is average difference in parameter between groups differing by 1 unit in X "holding covariates constant"
  - Statistical jargon: a "contrast" across the groups

35

35

Nonadaptive Randomization

Complete Randomization

Where am I going?

- The simplest form of randomization is independent randomization of each individual
- Within the context of a completely randomized design, we can explore its performance with respect to
  - Bias,
  - Face validity, and
  - Precision.

36

36

Randomization Strategies

- Complete randomization (CRD)
- Blocked randomization
  - Ensure balance after every *k* patients
  - Ensure closer adherence to randomization ratio
  - Undisclosed block sizes to prevent bias
- Stratified randomization
  - Separately within strata defined by strong risk factors
    - Lessens chance of randomization imbalance
  - Need to consider how many variables can be used
- Dynamic randomization
  - Adaptive randomization to achieve best balance on marginal distribution of covariates

37

37

Complete Randomization (CRD)

- With each accrued subject a (possibly biased) coin is tossed to determine which arm
  - Probability of treatment arm = $r / (r + 1)$
  - Independence of successive randomizations
- Issues
  - Bias
  - Face validity
  - Precision

38

38

CRD: Unbiased

- On average (across repeated experiments)
  - No correlation between treatment variable and other covariates
  - Individual type I errors come from samples in which other covariates are imbalanced

$$\beta_1 = \gamma_1 + \rho_{XW} \frac{\sigma_W}{\sigma_X} \gamma_2$$

39

39

Face Validity: Table 1

	Methotrexate Arm		Placebo Arm	
	n	Mean (SD; Min – Max)	n	Mean (SD; Min – Max)
Age (yrs)	132	50.4 (8.5; 32 - 69)	133	52.2 (8.5; 26 - 67)
Female	132	92.4%	133	92.5%
Pruritus score	116	7.7 (3.8; 4 - 16)	124	6.9 (3.8; 4 - 20)
Splenomegaly	131	8.4%	133	10.5%
Telangiectasia	132	4.6%	133	11.3%
Edema	132	6.1%	133	3.0%
Alkaline phosphatase	132	242.6 (145.9; 53 - 933)	133	245.0 (187.6; 66 - 1130)
ALT	131	54.5 (41.7; 12 - 202)	132	50.6 (41.4; 12 - 311)
Total bilirubin	132	0.7 (0.4; 0.1 - 2.7)	133	0.7 (0.4; 0.1 - 2.4)
Albumin	132	4.0 (0.3; 3.1 - 6.0)	133	4.0 (0.3; 3.0 - 4.8)
Prothrombin time INR	124	1.0 (0.1; 0.7 - 1.3)	132	1.0 (0.1; 0.7 - 1.3)
Mayo score	128	3.8 (0.8; 1.6 - 6.3)	133	3.9 (0.8; 1.6 - 6.1)
Avg stage	128	2.2 (0.9; 1.0 - 4.0)	128	2.3 (0.9; 1.0 - 4.0)
Avg fibrosis	128	1.2 (0.8; 0.0 - 3.0)	128	1.3 (0.9; 0.0 - 3.0)

40

40

CRD: Face Validity

- Table 1: Potential for imbalance in covariates
  - Depends on number of covariates and correlations among them
  - Probability of at least one "significant" imbalance

Number Displayed	Worst Case	Correlation 0.00	0.30	0.50	0.75	0.90
1	.050	.050	.050	.050	.050	.050
2	.100	.098	.095	.090	.081	.070
3	.150	.143	.137	.126	.104	.084
5	.250	.226	.208	.184	.138	.101
10	.500	.401	.353	.284	.193	.127
20	1.000	.642	.540	.420	.258	.154
50	1.000	.923	.806	.624	.353	.193

41

41

CRD: Face Validity

- Of course, statistical significance is not the issue

- "Conditional confounding"
  - How does unadjusted estimate compare to adjusted estimate?
  - Product of sample correlation between *X* and *W* and adjusted association between *Y* and *W*

$$\beta_1 = \gamma_1 + r_{XW} \frac{\sigma_W}{\sigma_X} \gamma_2$$

42

42

Demonstration of the Problem

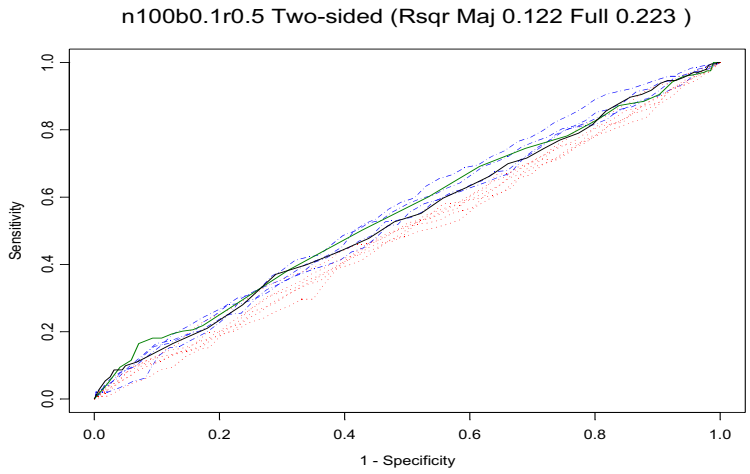
- Consider a CRD in presence of
  - 4 highly correlated predictors with larger importance
  - 6 independent predictors with smaller importance
  - No treatment effect
- Questions about unadjusted analysis
  - What is type I error? ➔ 0.025
  - What does imbalance in predictors tell us about type I error?
    - Sensitivity, specificity of imbalance in predictors under null hypothesis
    - Dependence on R^2 of measured covariates

43

43

Low Association: 2-sided

- ROC curve for covariate imbalance "explaining" statistical significance under the null

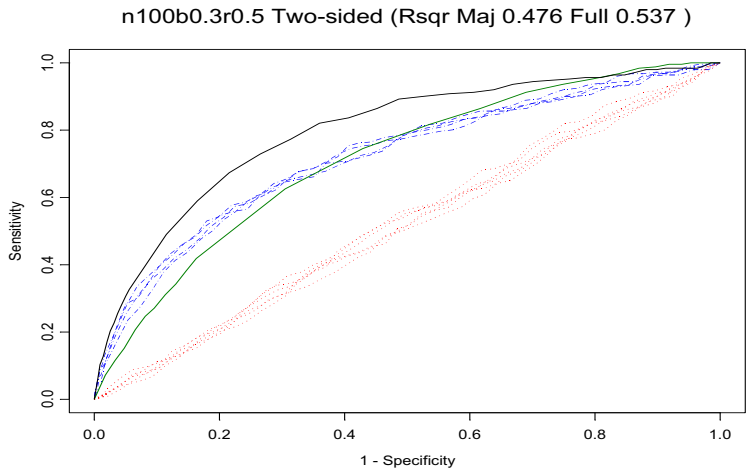


44

44

High Association: 2-sided

- ROC curve for covariate imbalance "explaining" statistical significance under the null

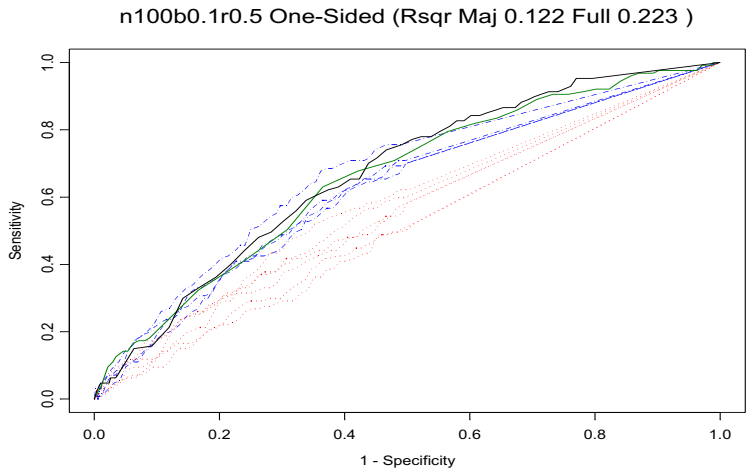


45

45

Low Association: 1-sided

- ROC curve for covariate imbalance "explaining" statistical significance under the null

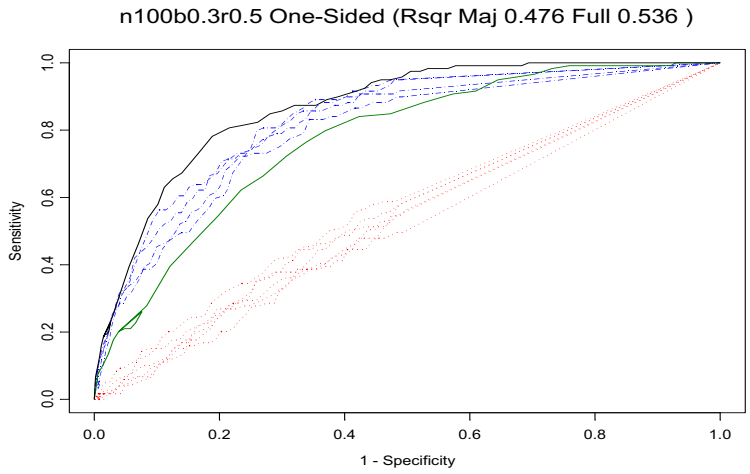


46

46

High Association: 1-sided

- ROC curve for covariate imbalance “explaining” statistical significance under the null



47

47

Face Validity

- Spurious results due to covariate imbalance
  - Unconditionally: Unbiased so no problem
  - Conditional on obtained randomization:
    - IF covariates are strongly predictive of outcome, then covariate imbalance is predictive of type I error
    - But need to consider that combined effect of other measured and unmeasured covariates may provide balance
- Ultimately, however, we need to have credible results
  - We do not always get to choose what others believe

48

48

Precision

- Impact of completely randomized design on precision of inference
  - Impact of imbalance in sample sizes
    - The number accrued to each arm is random
  - Impact of imbalance in covariates
    - "One statistician's mean is another statistician's variance"

49

49

Randomization Ratio

- Most efficient
  - When test statistics involve a sum, choose ratio equal to ratio of standard deviations
- Most ethical for patients on study
  - Assign more patients to best treatment
    - Many sponsors / patients presume new treatment
    - (Adaptive randomization: Play the winner)
- Most ethical for general patient population
  - Whatever is most efficient (generally not adaptive)
- Other goals
  - Attaining sufficient patients exposed to new treatment
  - Maintaining DSMB blind

50

50

Comment: Optimal *r* (Fixed *n*)

- Suppose we are constrained by maximal sample size $n = n_1 + n_2$
  - Smallest standard error in linear, logistic, Poisson regression models when

$$r = \frac{n_1}{n_2} = \frac{\sigma_1}{\sigma_2}$$

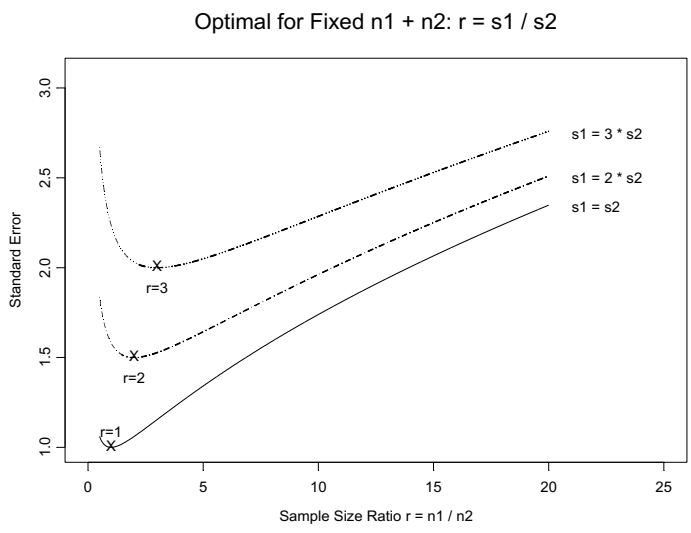
- In proportional hazards model smallest standard error when r = 1 ***in risk sets***

$$se(\hat{\theta}) = \sqrt{\frac{V}{n}} \quad \text{with} \quad V = \frac{(1+r)^2}{rPr(\delta_{ij}=1)} \qquad \hat{\theta} \sim \mathcal{N}\left(\theta, \frac{(1+r)^2}{r\,d}\right)$$

51

51

Comment: Optimal *r* (Fixed *n*)

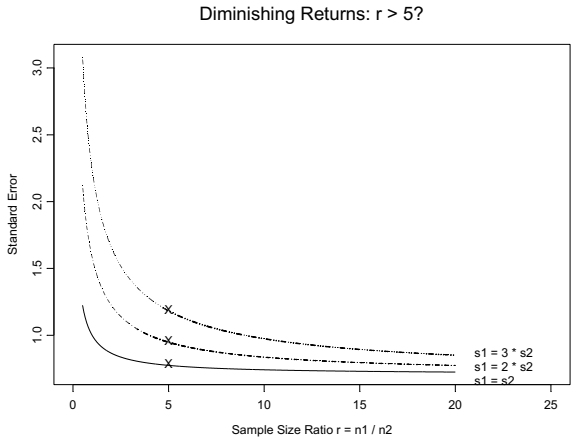


52

52

Comment: Diminishing Returns

- When we are unconstrained by maximal sample size we still hit a point of diminishing returns
  - Often quoted: *r = 5*



53

53

CRD: Efficiency Loss from Wrong Ratio

- CRD may not attain optimal ratio
  - Following table explores practical inefficiency
  - (True inefficiency is infinite due to possibility of no subjects randomized to one group)

N	r= 1	r= 2	r= 3	r= 5	r=10
20	1.0599	1.0652	1.0694	***	***
50	1.0213	1.0219	1.0229	1.0258	1.0282
100	1.0103	1.0104	1.0106	1.0111	1.0130
200	1.0051	1.0051	1.0051	1.0053	1.0056
500	1.0020	1.0020	1.0020	1.0020	1.0021
1000	1.0010	1.0010	1.0010	1.0010	1.0010

54

54

CRD: Efficiency Loss from Imbalance

- Covariates may be imbalanced across arms
  - Variability across replicated experiments increased if important predictor not controlled
  - Increased within group variance

Unadjusted Model $\left[se\left(\hat{\beta}_1\right)\right]^2 = \frac{Var\left(Y \mid X\right)}{nVar\left(X\right)}$

$$Var\left(Y \mid X\right) = \gamma_2^2 Var\left(W \mid X\right) + Var\left(Y \mid X, W\right)$$

55

55

CRD: Improved Performance

- If we adjust for important covariates, we will often gain precision
  - Face validity in Table 1 if readers recognize that adjustment accounts for any observed imbalance

- Caveats:
  - If covariate imbalance by arm, model misspecification can be an issue re conditional bias
  - If covariate imbalance by arm, lack of effect can be an issue re variance inflation
  - If adjustment not TOTALLY prespecified, “intent to cheat” analysis can be an issue
    - Not too much loss of precision from imperfect model

56

56

CRD: Linear Regr Continuous vs Dichotomized

Tx Eff	CRD – Continuous Adjust				CRD – Dichotomized Adjust			
	SE Slope		Power		SE Slope		Power	
	Unadj	Adj	Unadj	Adj	Unadj	Adj	Unadj	Adj
0.0	.281	.211	.026	.024	.284	.231	.023	.026
0.1	.278	.209	.053	.062	.284	.229	.045	.062
0.3	.279	.209	.178	.285	.287	.231	.184	.243
0.5	.281	.209	.423	.655	.279	.225	.409	.581
0.7	.279	.209	.696	.909	.281	.229	.699	.858

57

57

CRD: PH Regr Continuous vs Dichotomized

- Effect of W: log HR (HR) per SD(W), dichotomization at median
  - Number of events increases 5 – 10% with effect of W

	UnAdjusted			Adjusted Continuous			Adjusted Dichotomized		
W Eff	Tx HR Est	SE	Power	Tx HR Est	SE	Power	Tx HR Est	SE	Power
0.0 (1.00)	-0.201	0.191	0.178	-0.202	0.193	0.179	-0.201	0.192	0.176
0.3 (1.35)	-0.203	0.197	0.169	-0.209	0.200	0.185	-0.207	0.200	0.183
0.6 (1.82)	-0.181	0.195	0.152	-0.203	0.196	0.190	-0.196	0.195	0.168
0.9 (2.46)	-0.154	0.189	0.113	-0.199	0.188	0.160	-0.178	0.188	0.132
1.2 (3.32)	-0.148	0.188	0.116	-0.212	0.194	0.197	-0.185	0.190	0.143

	UnAdjusted			Adjusted Continuous			Adjusted Dichotomized		
W Eff	Tx HR Est	SE	Power	Tx HR Est	SE	Power	Tx HR Est	SE	Power
0.0 (1.00)	-0.503	0.199	0.700	-0.506	0.201	0.700	-0.505	0.200	0.700
0.3 (1.35)	-0.487	0.203	0.660	-0.504	0.207	0.694	-0.498	0.206	0.674
0.6 (1.82)	-0.449	0.209	0.614	-0.506	0.211	0.709	-0.486	0.209	0.672
0.9 (2.46)	-0.403	0.202	0.515	-0.507	0.206	0.693	-0.458	0.201	0.618
1.2 (3.32)	-0.358	0.195	0.433	-0.508	0.201	0.716	-0.432	0.199	0.575

	UnAdjusted			Adjusted Continuous			Adjusted Dichotomized		
W Eff	Tx HR Est	SE	Power	Tx HR Est	SE	Power	Tx HR Est	SE	Power
0.0 (1.00)	-0.818	0.219	0.972	-0.822	0.221	0.971	-0.821	0.220	0.971
0.3 (1.35)	-0.792	0.222	0.968	-0.816	0.227	0.967	-0.807	0.226	0.970
0.6 (1.82)	-0.733	0.215	0.935	-0.818	0.216	0.980	-0.784	0.217	0.962
0.9 (2.46)	-0.643	0.208	0.871	-0.807	0.213	0.977	-0.733	0.213	0.943
1.2 (3.32)	-0.560	0.198	0.774	-0.794	0.208	0.973	-0.680	0.204	0.912

58

58

Nonadaptive Randomization

Blocked Randomization

Where am I going?

- Blocking is sometimes used to ensure
  - Proper ratio of sample sizes across groups, and
  - Balance across arms over time

59

59

Randomization Strategies

- Complete randomization
- Blocked randomization
  - Ensure balance after every *k* patients
  - Ensure closer adherence to randomization ratio
  - Undisclosed block sizes to prevent bias
- Stratified randomization
  - Separately within strata defined by strong risk factors
    - Lessens chance of randomization imbalance
  - Need to consider how many variables can be used
- Dynamic randomization
  - Adaptive randomization to achieve best balance on marginal distribution of covariates

60

60

Issues with CRD

- Imbalance across arms in sample sizes
  - Not much of an issue with large sample sizes
  - Could be problematic with sequential sampling
    - Interim analyses of data early in the study
- Imbalance across arms in time trends
  - Outcome may be associated with time of accrual

61

61

Mechanisms Leading to Time Trends

- Patients accrued early may differ from those accrued later, because
  - Backlog of eligible patients
  - Startup of new clinical sites
  - Pressure to increase accrual
  - Secular trends in beliefs about intervention
    - (Made much worse if any interim results leak out)
  - Secular trends in diagnostic tools used for eligibility
  - Secular trends in ancillary treatments

62

62

Blocked Randomization

- Within every sequence of *k* patients, the ratio of treatment to control is exactly *r : 1*
  - Within each "block" ordering of treatments is random
- Important caveats:
  - Investigators must not know block size
    - Otherwise, decisions to enroll patients might be affected by knowledge of next assignment
  - Hence, often use "concealed blocks of varying sizes"
  - (Not really an issue in large multicenter RCT with central randomization)

63

63

Statistical Inference

- Impact on statistical inference relative to CRD
  - Bias properties unchanged
  - Face validity largely unchanged
    - We rarely report accrual patterns over time
  - Precision slightly improved due to achieving closer to desired randomization ratio
  - Precision could be improved if adjust for blocks as a random effect in analysis
    - This is rarely done, except in re-randomization test
      - Large number of small blocks, often with small variance of the random effects

64

64

Nonadaptive Randomization

Stratified Randomization

Where am I going?

- Stratified randomization is sometimes used to ensure proper ratio of sample sizes across subgroups defined by important covariates, thereby
  - Decreasing conditional bias,
  - Improving face validity, and
  - Possibly improving precision
- Major improvements in precision are gained only with adjustment for important stratification variables

65

65

Randomization Strategies

- Complete randomization
- Blocked randomization
  - Ensure balance after every *k* patients
  - Ensure closer adherence to randomization ratio
  - Undisclosed block sizes to prevent bias
- Stratified randomization
  - Separately within strata defined by strong risk factors
    - Lessens chance of randomization imbalance
  - Need to consider how many variables can be used
- Dynamic randomization
  - Adaptive randomization to achieve best balance on marginal distribution of covariates

66

66

Issues with CRD

- Imbalance across arms in covariate distribution
  - Loss of face validity
  - Conditional bias
  - Not much of an issue with large sample sizes
  - Could be problematic with sequential sampling
    - Interim analyses of data early in the study
  - Could be problematic with subgroup analyses
    - Possibility of very inefficient randomization ratio in small subgroups

67

67

Stratified Randomization

- Strata are defined based on values of important covariates
  - E.g., sex, age, disease severity, clinical site
- Within each stratum defined by a unique combination of stratification variables, CRD or blocked randomization
- Important caveats:
  - Number of strata is exponential in number of stratification variables
    - E.g., 4 two level stratification variables ➔ 16 strata

68

68

Statistical Inference

- Impact on statistical inference relative to CRD
  - Bias properties unchanged
  - Face validity improved for most important variables
  - Precision improved due to achieving closer to desired randomization ratio
  - Precision could be further improved if adjust for stratification variables in analysis
    - This should be done
      - Without adjustment for strata, may even lose power for some alternatives
    - Requires pre-specification of analysis model to avoid "intent to cheat" analysis

69

69

Lin Regr: CRD vs Orthogonal Randomization

	CRD – Continuous Adjust				Orthogonal Randomization			
Tx Eff	SE Slope		Power		SE Slope		Power	
	Unadj	Adj	Unadj	Adj	Unadj	Adj	Unadj	Adj
0.0	.281	.211	.026	.024	.206	.206	.005	.026
0.1	.278	.209	.053	.062	.208	.208	.013	.069
0.3	.279	.209	.178	.285	.205	.205	.115	.313
0.5	.281	.209	.423	.655	.205	.205	.403	.684
0.7	.279	.209	.696	.909	.205	.205	.759	.924

70

70

PH Regr: CRD vs Orthogonal Randomization

- Effect of W: log HR (HR) per SD(W), dichotomization at median
  - ORTH 1 is each time different design, ORTH 2 is single design
  - True SE (displayed) sometimes less than est SE (stays constant)

CRD	UnAdjusted			Adjusted Continuous			Adjusted Dichotomized		
W Eff	Tx HR Est	SE	Power	Tx HR Est	SE	Power	Tx HR Est	SE	Power
0.0 (1.00)	-0.503	0.199	0.700	-0.506	0.201	0.700	-0.505	0.200	0.700
0.3 (1.35)	-0.487	0.203	0.660	-0.504	0.207	0.694	-0.498	0.206	0.674
0.6 (1.82)	-0.449	0.209	0.614	-0.506	0.211	0.709	-0.486	0.209	0.672
0.9 (2.46)	-0.403	0.202	0.515	-0.507	0.206	0.693	-0.458	0.201	0.618
1.2 (3.32)	-0.358	0.195	0.433	-0.508	0.201	0.716	-0.432	0.199	0.575

ORTH 1	UnAdjusted			Adjusted Continuous			Adjusted Dichotomized		
W Eff	Tx HR Est	SE	Power	Tx HR Est	SE	Power	Tx HR Est	SE	Power
0.0 (1.00)	-0.505	0.204	0.678	-0.506	0.205	0.681	-0.506	0.204	0.683
0.3 (1.35)	-0.505	0.204	0.697	-0.523	0.210	0.724	-0.518	0.208	0.714
0.6 (1.82)	-0.437	0.187	0.581	-0.492	0.204	0.666	-0.470	0.196	0.634
0.9 (2.46)	-0.399	0.175	0.512	-0.504	0.206	0.710	-0.457	0.195	0.632
1.2 (3.32)	-0.358	0.156	0.419	-0.511	0.198	0.730	-0.438	0.180	0.606

ORTH 2	UnAdjusted			Adjusted Continuous			Adjusted Dichotomized		
W Eff	Tx HR Est	SE	Power	Tx HR Est	SE	Power	Tx HR Est	SE	Power
0.0 (1.00)	-0.495	0.203	0.687	-0.497	0.204	0.686	-0.497	0.203	0.683
0.3 (1.35)	-0.491	0.190	0.690	-0.508	0.194	0.718	-0.503	0.192	0.709
0.6 (1.82)	-0.442	0.186	0.592	-0.497	0.204	0.695	-0.476	0.197	0.653
0.9 (2.46)	-0.400	0.168	0.520	-0.506	0.201	0.693	-0.459	0.191	0.641
1.2 (3.32)	-0.345	0.157	0.393	-0.507	0.204	0.727	-0.429	0.185	0.581

1

71

Advantages

- Additional advantages of stratification
  - Balance within clinical center
    - Especially if quality control issues
  - Balance for interim analyses
  - Balance for subgroup analyses

72

72

Adaptive Randomization

Covariate Adaptive Randomization

Where am I going?

- Stratified randomizations has drawbacks in the presence of sparse data
- Some authors have described dynamic randomization processes that will allow balancing on more covariates

73

73

Issues with Stratified Analyses

- The need to stratify on all combinations of variables
  - Good news:
    - Balances on interactions as well as main effects
  - Bad news:
    - Effect of interactions might be quite small
    - Really only need to adjust on "counterfactual" outcome based on linear combination of all covariates

74

74

Dynamic Randomization

- Subjects are assigned to the treatment arm that will achieve best balance
- "Minimization": minimize the difference between the distribution of covariate effects between arms
  - Define a "distance" between arms for covariate vectors
  - Probability of assignment depends upon arm that would provide smallest difference

75

75

Distance Between Arms

- Two arms are "distant" based on one of:
  - Randomization ratio very different from *r : 1* in some stratum
  - Summary measure of distribution of $(W_{i1},\ldots,W_{ip})$ differs
    - Mean, median, variance, …
  - Distribution of $(W_{i1},\ldots,W_{ip})$ differs
  - Contribution of covariates to the outcome differs

76

76

Conditional Confounding

$$\text{Unadjusted}: \quad g\left[\theta \,|\, X_i\right] = \beta_0 + \beta_1 \times X_i$$

$$\text{Adjusted} \quad : \; g\left[\theta \,|\, X_i, \vec{W}_i\right] = \gamma_0 + \gamma_1 \times X_i + \vec{W}_i^T \vec{\delta}$$

$$\text{Unadjusted}: \quad g\left[\theta \,|\, \mathbf{X}\right] = \mathbf{X}\vec{\beta}$$

$$\text{Adjusted} \quad : \; g\left[\theta \,|\, \mathbf{X}, \mathbf{W}\right] = \mathbf{X}\vec{\gamma} + \mathbf{W}\vec{\delta}$$

$$\vec{\beta} = \vec{\gamma} + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\vec{\delta}$$

77

77

Conditional Confounding

$$g\left[\theta \,|\, \mathbf{X}\right] = \mathbf{X}\vec{\beta} \qquad g\left[\theta \,|\, \mathbf{X}, \mathbf{W}\right] = \mathbf{X}\vec{\gamma} + \mathbf{W}\vec{\delta}$$

$$\vec{\beta} = \vec{\gamma} + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\vec{\delta}$$

$$\beta_1 = \gamma_1 + \sum_{j=1}^{p}\left(\overline{W}_{1j\bullet} - \overline{W}_{0j\bullet}\right)\delta_j$$

$$\overline{W}_{kj\bullet} = \frac{1}{n_k}\sum_{i=1}^{n} W_{ij} 1_{[X_i = k]}$$

78

78

Distance Metrics

Based on contribution to confounding :

$$d\left(\vec{X},\mathbf{W}\right)=\left|\sum_{j=1}^{p}\left(\overline{W}_{1j\bullet}-\overline{W}_{0j\bullet}\right)\delta_j\right|$$

Weighted distance between standardized means :

$$d\left(\vec{X},\mathbf{W}\right)=\sum_{j=1}^{p}c_j\left|\frac{\overline{W}_{1j\bullet}-\overline{W}_{0j\bullet}}{SD\left(W_j\right)}\right|^{\lambda}$$

Weighted imbalance in n across strata $\Omega_1,\ldots,\Omega_s$:

$$d\left(\vec{X},\mathbf{W}\right)=\sum_{s=1}^{S}c_s\left|\sum_{i=1}^{n}1_{[X_i=1]}1_{[\vec{W}_i\in\Omega_s]}-\sum_{i=1}^{n}1_{[X_i=0]}1_{[\vec{W}_i\in\Omega_s]}\right|^{\lambda}$$

79

79

Implication

- Spurious associations will be minimized if means of important predictors are balanced across treatment arms
  - The greater the value of δ_j the more important it is for the <u>means</u> of the j-th covariate to be equal
    - (Presumes linear model reasonable approximation)
  - We could use estimates of the of δ_j's to define the distance between the arms (or just balance means)
- Balancing group sizes across covariates will tend to have means balanced by randomization
  - Group sizes within strata may matter for subgroup analyses

80

80

Probability of Assignment

- Subjects are assigned to the treatment arm that will achieve best balance
  - When i-th patient accrued, compute a randomization probability

$$\Delta_i = d\left(\vec{X}, \mathbf{W} \mid X_i = 1\right) - d\left(\vec{X}, \mathbf{W} \mid X_i = 0\right)$$

$$\pi_i = \Pr\left(X_i = 1\right) = f\left(\Delta_i\right)$$

- $0 \leq \pi_i \leq 1$
- Larger values of Δ_i ➔ smaller values of π_i
- Probably best to avoid $\pi_i = 0$ and $\pi_i = 1$

81

81

Inference: Population Model

- Impact on statistical inference relative to CRD
  - Bias properties unchanged
  - Face validity improved for most important variables
  - Precision improved due to achieving closer to desired randomization ratio
  - Precision could be further improved if adjust for stratification variables in analysis for population model
    - This should be done
    - Requires pre-specification of analysis model to avoid “intent to cheat” analysis

82

82

Inference: Randomization Model

- Proschan, Brittain, Kammerman, 2010: Precision could be greatly hampered if you analyze under randomization hypothesis
- Alternative randomization schemes may be quite restrictive, especially under unequal randomization
  - Suppose sequential allocation
    - Randomization P value is identically 1 (or 0.5?)
  - If dynamic randomization has $\pi_i = 0$ or $\pi_i = 1$ too often, range of randomization P values is greatly restricted
- Also: Statistical analysis can be quite involved

83

83

Advantages / Disadvantages

- Advantages
  - Typically improved face validity
  - Can handle an arbitrary number of covariates
    - Depending on distance metric
- Disadvantages
  - Logistically more involved
  - Decreased credibility if too deterministic
    - Approaches sequential allocation
  - Some analytic strategies more complex
  - Does not necessarily facilitate subgroup analyses
    - Unless distance metric chosen carefully

84

84

Adaptive Randomization

Response Adaptive Randomization

Where am I going?

- Some authors have described dynamic randomization processes that attempt to minimize exposure of patients to harmful treatments

85

85

Ethics

- Clinical trials are experiments in human volunteers
- Individual ethics
  - Patients on trial: Avoid continued administration of inferior treatment
  - Patients not yet on trial: Avoid starting inferior treatment
- Group ethics
  - Facilitate rapid adoption of new beneficial treatments
  - Avoid prolonging study of ineffective treatments

86

86

Solutions

- Most commonly used
  - Sequential sampling
    - Interim analyses of data
    - Terminate trials when credible decisions can be made
- Also proposed
  - Response adaptive randomization
    - Change randomization probabilities as evidence accumulates that one treatment might be best
    - "Play the winner"

87

87

Play the Winner: Urn Model

- Begin with *k* white balls and *k* black balls in an urn
- Upon accrual of a patient draw a ball from urn
  - White ➔ control; black ➔ treatment
- Observe outcome
  - If outcome is good, return *m+1* balls of same color as withdrawn
  - If outcome is bad, return 1 ball of same color as withdrawn and *m* balls of opposite color

88

88

Bayesian Methods

- An explicit Bayesian approach could to dynamic randomization could base the randomization ration on the current posterior probability that one treatment is superior
  - Ultimately, that posterior probability is based on the number of good outcomes on each treatment
- Advantage of using Bayesian posterior probability
  - Can easily handle continuous outcomes
  - Can easily handle continuous randomization probabilities

89

89

Analytic Issues

- Treatment of successive patients is not independent of previous patients treatment and results
  - Possible bias in accrual of future patients
- Conditionally biased estimates of treatment effect in arm with lower sample sizes
  - Bad early results tend to preclude regression to mean
- Randomization hypothesis can lead to quite unconvincing results

90

90

Example: ECMO Study

- Randomized clinical trial of extracorporeal membrane oxygenation in newborns
  - Randomized PTW design with k=1
- Data:
  - First patient on ECMO survived
  - Next patient on control died
  - Next 9 patients on ECMO survived
- Inference (Begg, 1990)
  - P value of 0.001, 0.051, 0.083, 0.28, 0.62?

91

91

Comments

- This experience has tempered enthusiasm for randomized PTW
  - Interestingly, follow-up studies had 67% survival on conventional therapy
- I believe there can be times that this will work, but
  - There needs to be a clear dilemma re individual ethics
  - There will tend to be decreased group ethics
  - It takes a lot of planning in order to obtain results that will be sufficiently credible
    - Assuming your conclusion will not cut it

92

92

More Recent Adaptive Designs

- Some recent efforts have been directed toward more rapid identification of treatments worthy of Phase 3 study
  - E.g., I-SPY2
- Consider a 5 arm RCT of 4 experimental therapies and placebo
  - Initially randomize 1:1:1:1:1
- At an interim analysis, accelerate accrual to most promising arms
  - For instance, 4:0:0:0:1
  - (Statistic used to make decision is not crucial)
- Complete investigation of that most promising arm (and progress to Phase 3 as appropriate),
  - Resume investigation to the other arms

93

93