Biost 524: Design of Medical Studies

Lecture 4:
Comparison Groups; Randomization; Blinding

Susanne J. May, Ph.D. / Scott S. Emerson, M.D.,Ph.D.
Associate Professor / Professor of Biostatistics
University of Washington

April 20, 2011

1

Comparison Groups

Options

Where am I going?

- Having a comparison group is important when
  - deciding whether a proposed treatment is effective, and
  - deciding among the alternatives when treating a single patient

2

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)

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

Historical Controls

- An attempt to make more efficient use of limited research resources
- 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

5

Sample Size: Single Arm Study

- Sample size requirements in a single arm study to detect a mean outcome greater than μ_0

$$n = \frac{(z_{1-\alpha/2} + z_\beta)^2 \sigma^2}{(\mu_1 - \mu_0)^2}$$

6

Sample Size: Two Arm Study

- Sample size requirements on experimental arm in a two arm study to detect a mean outcome greater than μ_0
  - $n_1 = r \times n_0$ with r the ratio between sample sizes

$$n_1 = \frac{(z_{1-\alpha/2} + z_\beta)^2 \left(1 + n_1/n_0\right) \sigma^2}{(\mu_1 - \mu_0)^2}$$

7

Sample Size: Historical Controls

- Sample size requirements on experimental arm when using historical controls in a study to detect a mean outcome greater than μ_0
  - n_0 historical controls are presumably already available

$$n_1 = \frac{(z_{1-\alpha/2} + z_\beta)^2 \left(1 + n_1/n_0\right) \sigma^2}{(\mu_1 - \mu_0)^2}$$

8

Use of Historical Controls

- Compared to a two arm study of a new treatment and a historical treatment, use of historical data can save time and money
  - Use of a historical control sample obviates the need for one arm: thus only half the subjects when 1:1 randomization.
  - Using the estimates from a historical clinical trial as if they were known treatment effects decreases sample size requirements even further:
    - Only one-fourth the number of subjects are required
    - But pretending that we have an infinite number of relevant historical controls

9

Use of Historical Controls

- 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
- Pocock (*J Chronic Disease*, 1976) described conditions for acceptability of historical control group

10

Pocock Conditions - 1

- Such a group must have received a precisely defined standard treatment
  - relevance of standard treatment must remain
  - measurement of treatment parameters must be the same
  - ancillary treatments must not have changed

11

Pocock Conditions - 2

- Group must have been a part of a recent clinical study containing the same requirements for patient eligibility
  - measurement methods used in eligibility must be the same
  - clinical trial setting must have same selection pressures on patient participation

12

Pocock Conditions - 3

- Methods of treatment evaluation must be the same
  - same criteria (schedule) for performing evaluations
  - same criteria for judging outcomes

13

Pocock Conditions - 4

- Distributions of important patient characteristics should be comparable
  - same univariate distributions of risk factors (within range dictated by eligibility criteria)
  - same correlations among risk factors
  - must hold for both measured and unmeasured risk factors of
    - disease,
    - (propensity for) adverse outcomes,
    - and competing risks

14

Pocock Conditions - 5

- Previous study must have been performed in the same organization with largely the same clinical investigators
  - must control any subjective aspects of definition of eligibility, treatments, outcome
  - must control for unique patient populations due to location and/or referral patterns

15

Pocock Conditions - 6

- There must be no other indications leading one to expect differing results

16

Additional Criterion

- The analysis should reflect the variability in the original data, not just the estimates of treatment effect
  - It is “cheating” to pretend there was no variability in assessing the outcome from the historical comparison group.
  - Ideally: use the exact distribution of the covariates
    - Nonlinearities of effects of covariates on outcome and interactions among the covariates might alter the inference

17

Proposed Remedies

- Attempts to circumvent some of these requirements using statistical methods
  - Clearly, the above conditions are rarely, if ever, satisfied.
  - Attempts have been made to use statistical models to adjust for differences between the historical control group and a current treatment group.
    - Adjustment for covariates
    - Propensity score analysis

18

Adjustment for Covariates

- Analysis with adjustment for confounding due to dissimilarities between treatment groups
  - Adjust for important predictors of treatment outcome
  - E.g., analyze treatment effect in a regression model including indicator of treatment
  - include as covariates those prognostic variables that differ between the groups

19

Propensity Score Analyses

- Propensity score analyses
  - Attempts to mimic randomization; does not worry about prognostic capability for outcome
    - Confounding = association between covariate and treatment AND association between covariate and outcome
  - Creates a “propensity score” measuring the propensity for an individual with specific covariates to be in the new treatment group
  - Perform an analysis adjusting for propensity scores
    - In each stratum, there is no association between covariate and treatment

20

Drawbacks

- Both approaches 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.

21

Problem with Time Trends

- Furthermore, progress in diagnostic methods and therapeutic strategies means that few measurements made in the past are exactly comparable to those made now
  - Laboratory and imaging techniques lead to improved diagnosis and staging of disease
    - E.g., earlier diagnosis of disease may lead to perceived better survival
    - E.g., detection of metastases at earlier stages causes trends toward milder disease being diagnosed as Stage IV
  - Supportive measures may improve outcomes

22

Final Comments re Historical Controls

- The use of historical controls from previous clinical trials would thus appear highly problematic
  - The situation is only worse if one tries to use data from cohort studies or other observational data bases
  - Such data bases are well suited for hypothesis generation and feasibility studies, but do not at all provide comparability to a clinical trial setting

23

Final Comments re Historical Controls

- Nevertheless, when strong treatment effects are expected and when little has changed in the disease setting, the use of historical controls may be the only ethical option
  - Sometimes the window of opportunity for a randomized trial is extremely short
    - Early feasibility studies might show such promising results that preclude equipoise
    - (Some authors therefore suggest randomizing at every stage of investigation)

24

Example: ECMO Study

- Randomized clinical trial of extracorporeal membrane oxygenation in newborns
  - Randomized Play The Winner (PTW) design in which randomization ratio changes to favor more "successful" therapy
- Data:
  - First patient on ECMO survived
  - Next patient on control died
  - Next 9 patients on ECMO survived

25

Comments

- Arguments for credibility of results
  - Prior history suggested 90% mortality under standard of care
- Inference (Begg, 1990)
  - P value of 0.001, 0.051, 0.083, 0.28, 0.62?
- This experience has tempered enthusiasm for randomized PTW
  - Interestingly, follow-up studies had 67% survival on conventional therapy

26

Internal Controls

- Each subject serves as his/her own control
  - Different treatments at different times
  - Different treatments for different parts of body
    - eye diseases, skin diseases
- N.B.: This does not include "pre-post" designs looking at the change from baseline in a single arm study
  - These would be uncontrolled experiments

27

Concurrent Control Groups

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

28

Blinding

Options

Where am I going?

- Participant and investigator biases can be (and have been) a major source of bias in RCT

29

Definitions

- 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 blinded

30

Goals

- Blinding can serve to
  - Minimize "placebo effect": A participant being treated does better than one not being treated, irrespective of the actual treatment
    - This should be distinguished from secular trends in outcome that might happen over time
    - To detect a placebo effect, you compare a group that is unknowingly receiving a placebo to a group that is receiving nothing
  - Minimize investigator bias in assessing
    - adverse events
    - treatment outcomes

31

Logistical Issues

- Blinding is not always possible
  - Placebo not always possible to be identical in appearance
    - weight of fiber, hardness of calcium,
  - Side effects of treatment may be noticeable
    - skin discoloration with beta-carotene
  - Burden of treatment may not be ethical
    - surgery, hospitalizations

32

Blinded Evaluation

- When blinding of participants and investigators is not possible, blinded evaluation may be
  - Must still ensure similar schedule of assessment
    - side effects might lead to more frequent monitoring
  - Competing risks (e.g., death from other causes) still a problem

33

Issues

- Issues that must be addressed with blinded experiments
  - Appearance of treatments
  - Dosage, administration schedules
  - Coding and dispensing treatments
  - When and how to unblind
    - Emergent situations
    - Only unblind when treatment of toxicities differs between therapies
  - Assessing how well the blind was maintained

34

When Blinding is Unnecessary

- Blinding less of an issue with harder endpoints
  - The more objective the measurement of outcome, the less important blinding is to the scientific credibility of a clinical trial.
  - (Of course, the ideal is a blinded experiment with solidly objective endpoints.)

35

Why Randomize?

Statistical Role of Variables in Analysis

Where am I going?

- Ultimately a RCT is designed to compare outcomes across groups in a statistical analysis
- It is useful to review the components of a statistical analysis model in order to
  - develop a standard nomenclature and
  - discuss the goals and impact of randomization.

36

Second Statistical Refinement

- The group receiving the treatment will tend to have outcome measurements that are

higher than,	an absolute standard, or
lower than, or	measurements in an otherwise comparable group (that did not receive the treatment)
about the same as	

37

Ex: Smoking Effect on FEV

- Scientific question
  - Does smoking lead to lower lung function in kids?
- Study design
  - 654 healthy children
  - Measure smoking by self report
  - Measure lung function by FEV
    - Forced expiratory volume: maximum volume of air that can be exhaled in 1 second

38

GM: Unadjusted Interpretation

- Smoking effect
  - Geometric mean of FEV is 10.8% higher in smokers than in nonsmokers
    - 95% CI: 4.1% to 17.9% higher
    - These results are atypical of what we might expect with no true difference between groups: P = 0.001

39

Age, Ht Adjusted Interpretation

- Smoking effect
  - Geometric mean of FEV is 5.2% lower in smokers than in nonsmokers of the same age and height
    - 95% CI: 9.6% to 0.6% lower
    - These results are atypical of what we might expect with no true difference between groups of the same age and height: P = 0.027

40

Ex: Take Home Message

- Our scientific question was not

  Is there a difference between smokers' and nonsmokers' average FEV?

- But rather ("personalized medicine"?)

  Do smokers average lower FEV than otherwise comparable nonsmokers?

41

Problem

- While our adjusted regression analysis appears to come closer to answering our scientific question, it is still an observational data analysis
  - Our statistical "adjustment" is only an approximation
  - We had to "borrow" information according to a model
  - We can not infer cause and effect
- In this case, this is probably the best we can do
  - It is unethical to randomize children to smoke
- But our ideal will be to gain information faster
  - Randomized interventional experiments

42

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

43

Unadjusted, Adjusted Analyses

- In blinded RCT, we avoid confounding on average
  - But we do worry about imbalances in any single trial
- Confounding typically produces a difference between unadjusted and adjusted analyses, but those symptoms are not proof of confounding
  - Such a difference can occur times when there is no confounding
    - "Precision" variables in logistic, PH regression
    - Complicated causal pathways

44

Precision

- Sometimes we choose the exact scientific question to be answered on the basis of which question can be answered most precisely
  - In general, questions can be answered more precisely if the within group distribution is less variable
    - Comparing groups that are similar with respect to other important risk factors decreases variability

45

Next

Controlling Variation

- In a two sample comparison of means, we might control some variable in order to decrease the within group variability
  - Restrict population sampled (subgroup analysis)
  - Standardize ancillary treatments
  - Standardize measurement procedure
- In RCT, we can also consider
  - Stratified randomization to ensure balance across arms
  - Adjusted analyses to gain precision

46

Why Randomize?

Four Important Questions of Regression

Where am I going?

- The fundamental statistical distinctions between unadjusted and adjusted regression models are central to the goals of randomization
- We thus want to be able to consider the relationships between
  - unadjusted and adjusted parameters, and
  - the standard errors of the two parameter estimates.

47

Adjustment for Covariates

- We “adjust” for other covariates
  - Define groups according to
    - Predictor of interest, and
    - Other covariates
  - Compare the distribution of response across groups which
    - differ with respect to the Predictor of Interest, but
    - are the same with respect to the other covariates
      - “holding other variables constant”

48

Unadjusted vs Adjusted Models

- Adjustment for covariates changes the scientific question
  - Unadjusted models
    - Slope compares parameters across groups differing by 1 unit in the modeled predictor
      - Groups may also differ with respect to other variables
  - Adjusted models
    - Slope compares parameters across groups differing by 1 unit in the modeled predictor but similar with respect to other modeled covariates

49

Interpretation of Slopes

- Difference in interpretation of slopes

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

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

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

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

50

Comparing models

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

Adjusted $\quad g\left[\theta \,|\, X_i, W_i\right] = \gamma_0 + \gamma_1 \times X_i + \gamma_2 \times W_i$

Science: When is $\gamma_1 = \beta_1$?

When is $\hat{\gamma}_1 = \hat{\beta}_1$?

Statistics: When is $se(\hat{\gamma}_1) = se(\hat{\beta}_1)$?

When is $\hat{se}(\hat{\gamma}_1) = \hat{se}(\hat{\beta}_1)$?

51

Linear Regression

- Difference in interpretation of slopes

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

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

$$\text{Adjusted Model}: \; E\left[Y_i \,|\, X_i, W_i\right] = \gamma_0 + \gamma_1 \times X_i + \gamma_2 \times W_i$$

  - γ_1 = Diff in mean Y for groups differing by 1 unit in X, but agreeing in their values of W

52

Relationships: True Slopes

- The slope of the unadjusted model will tend to be

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

- Hence, true adjusted and unadjusted slopes for X are estimating the same quantity only if
  - $\rho_{XW} = 0$ (X and W are truly uncorrelated), OR
  - $\gamma_2 = 0$ (no association between W and Y after adjusting for X)

53

Linear Regression Results

- Adjusted and unadjusted true parameters agree IF
  - **mean** of W is equal across groups on average
- Adjusted and unadjusted estimated slopes agree IF
  - sample correlation between X and W is exactly 0
  - (when X is binary ➔ sample **means** of W are equal)
- Adjusted and unadjusted true SE agree IF
  - sample correlation between X and W is exactly 0, AND
  - W does not truly predict outcome
- Adjusted and unadjusted estimated SE do not agree

54

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 predictors, or
  - Adjust for stratification variables
- Subgroup analyses
  - If effect modification is concern

55

Nonadaptive Randomization

Randomization

Where am I going?

- The goal of a RCT is to assess cause and effect
- Randomization is the tool that allows this, but only for the scientific and statistical hypotheses that are based on randomization

56

Treatment of Variables

- Measure and compare distribution across groups
  - Response variable in regression
- Vary systematically (intervention)
- Control at a single level (fixed effects)
- Control at multiple levels (fixed or random effects)
  - Stratified (blocked) randomization
- Measure and adjust (fixed or random effects)
- Treat as "error"

57

Predictor of Interest

- The predictor of interest is varied systematically
  - *r* subjects on experimental treatment : 1 control

58

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)

59

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.

60

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

61

Randomization Strategies

- Complete randomization (CRD)
- Blocked randomization
  - Ensure balance after every *k* patients
- Stratified randomization
  - Separately within strata defined by strong risk factors
    - Lessens chance of randomization imbalance
- Dynamic randomization
  - Adaptive randomization to achieve best balance on marginal distribution of covariates
- Response adaptive randomization
  - E.g., “play the winner”

62

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.

63

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
  - Unbiased (on average)
  - Face validity: Imbalances may occur
  - Precision: May not achieve ratio, balance every time

64

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

65

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

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

66

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)

67

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

68

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*

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

69

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

70

CRD: Improved Performance

- If we adjust for important covariates, we will often gain precision (by removing "conditional confounding")
  - 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

71

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

72

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

73

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"

74

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

75

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

76

Advantages

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

77

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

78

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

79

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

80

Conditional Confounding

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

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

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

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

81

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
    - e.g., mean
  - Distribution of $(W_{i1},\ldots,W_{ip})$ differs
  - Contribution of covariates to the outcome differs

82

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

83

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

84

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

85

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"

86

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

87

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

88

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

89

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?

90

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

91

Randomization

Analytic Models

Where am I going?

- Randomization serves as the basis for ascribing cause and effect
- However, to realize this we must consider the statistical foundations for inference, which include
  - Population model
  - Randomization model

92

Analysis: 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

93

Analysis: 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

94

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
  - But strong null not in keeping with scientific method
    - Assumptions more detailed than primary question
      - Primary question usually about first moment
      - Semiparametric assumptions are about all moments
    - Consider bone marrow transplantation

95

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)

96

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"

97

Randomization

Overall Recommendations

Where am I going?

- So what do I really recommend?

98

Overall Recommendations

- Blind as much as possible
- Randomize
  - 1:1 randomization unless <u>really</u> good reason
- Stratify on
  - Clinical center
  - No more than 4 important variables (including center)
- Use hidden blocks of varying sizes
  - Block sizes 4, 6, or 8 in 1:1 randomization
- Prespecify population analysis model including covariates that are thought *a priori* to be the most highly predictive
  - (Less of an issue with logistic regression)

99