# Covariate Adjustments in Clinical Trials: Are They Worth the Added Complexity?

**Theodore Karrison**

**Society for Clinical Trials**

**May, 2009**

# Outline

- Inference in RCTs
- Reasons for covariate adjustment
- Examples

  REMIS, MESO, PBC, *NOVA*
- Problems with adjustment
- Solutions
- NOVA trial
- Conclusions

# Inference in Randomized Clinical Trials

"RCTs have become the standard by which the comparison of treatments can be judged. Randomization plays a central role because it eliminates selection bias in the treatment assignment, it tends to produce control and treatment groups that are comparable with respect to risk factors, and it provides the framework for the validity of the statistical inference."

Beach and Meier (CCT, 1989)

The (unadjusted) p-value informs us as to whether the observed difference is likely to be due to chance:

"The exact meaning of the familiar abbreviation 'P<0.05' is thus: the patients in one group have fared better than the patients in the other. If there is no difference between the medical effects of the two treatments and the only cause of differences between the treatment groups is the chance allocation of more good-prognosis patients to one group than to the other, then the chance of one treatment group faring at least this much better than the other groups would be less than 0.05, i.e., less than a 1 in 20 chance."

Peto, Pike, et al. (Br. J. Cancer, 1976)

# Reasons for Covariate Adjustment

So why adjust?

- Remove the effects of *chance* imbalances on the treatment comparison ("random bias")

  Alternatives: minimization, covariate-adapted randomization

- Improve precision, i.e. reduce SE of treatment comparison

- For non-linear models, provide a more "subject-specific" measure of the treatment effect (Hauck, Anderson, Marcus, CCT, 1998)

# Examples

- REMIS

  Recurrent Miscarriage Trial

- MESO

  Randomized phase II trial in mesothelioma

- PBC

  Mayo Clinic trial of D-penicillamine for primary biliary cirrhosis

- *NOVA*

  *UC trial of nitric oxide in premature infants*

# Problems with Adjustment

"Data dredging" can move the p-value around quite a bit

Analyst can pick the adjusted result that "accentuates the estimate and/or statistical significance of the treatment effect"

Pocock et. al (SIM, 2002)

P-value shopping . . .

# Beach and Meier (1989):

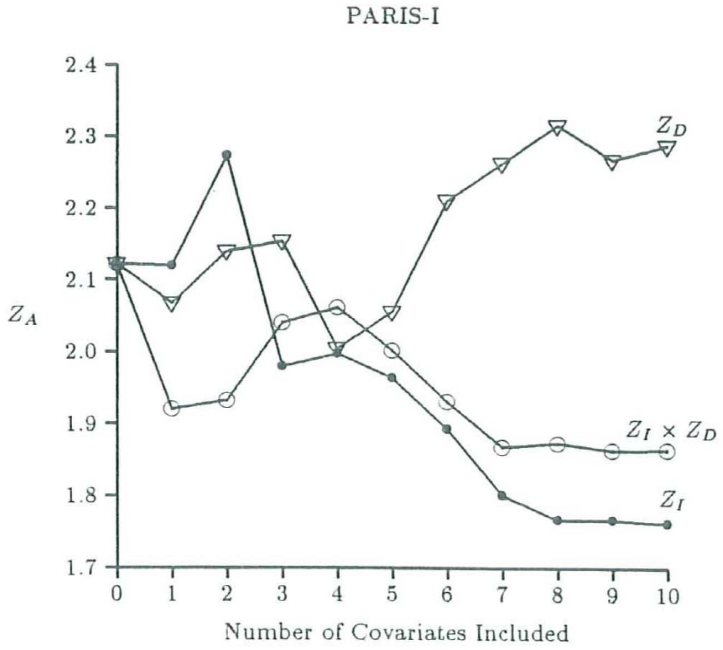



**Figure 4** PARIS–I: $Z_A$ versus number of covariates selected for three different methods of covariate selection.

Classic ANCOVA theory for linear models:

$$\boldsymbol{Z}_1 = \begin{cases} 1, \text{ TREATMENT} \\ 0, \text{ CONTROL} \end{cases} \qquad \boldsymbol{X}_2 = \text{covariate}$$

$$E(Y \mid Z_1) = a^* + b_1^* Z_1 \qquad Var(Y \mid Z_1) = \sigma^2_{Y.1}$$

$$E(Y \mid Z_1, X_2) = a + b_1 Z_1 + b_2 X_2 \qquad Var(Y \mid Z_1, X_2) = \sigma^2_{Y.12}$$

$$\boldsymbol{EFF} = \frac{\boldsymbol{Var}(\hat{\boldsymbol{b}}_1^*)}{\boldsymbol{Var}(\hat{\boldsymbol{b}}_1)} = \frac{1-\rho_{12}^2}{1-\rho_{\boldsymbol{Y}2.1}^2}$$

In RCT, $\rho_{12} = 0$ therefore

$$\boldsymbol{EFF} = \frac{1}{1-\rho_{\boldsymbol{Y}2.1}^2} \geq 1$$

Surprise: For common non-linear models, when analyzed in the usual way, adding covariates does *not* increase the precision of the treatment comparison!

REMIS trial (Ober et al., Lancet, 1999)

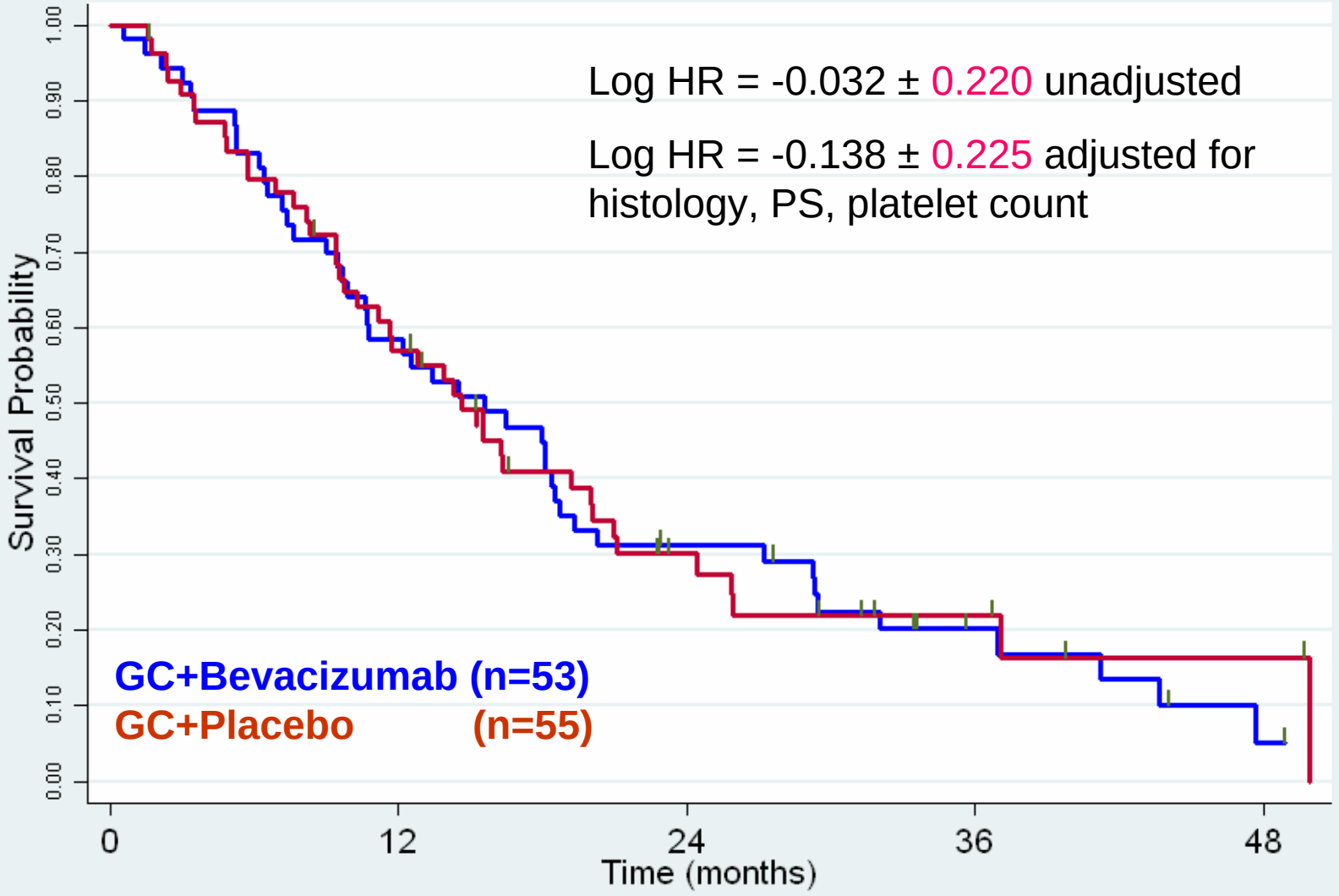
| Pregnancy outcome | Treatment group | Control group |
|---|---|---|
| All randomised patients | | |
| Total | 86 | 85 |
| Success | 31 (36%) | 41 (48%) |
| Failure | 55 (64%) | 44 (52%) |

Logistic regression model (estimate ± SE):

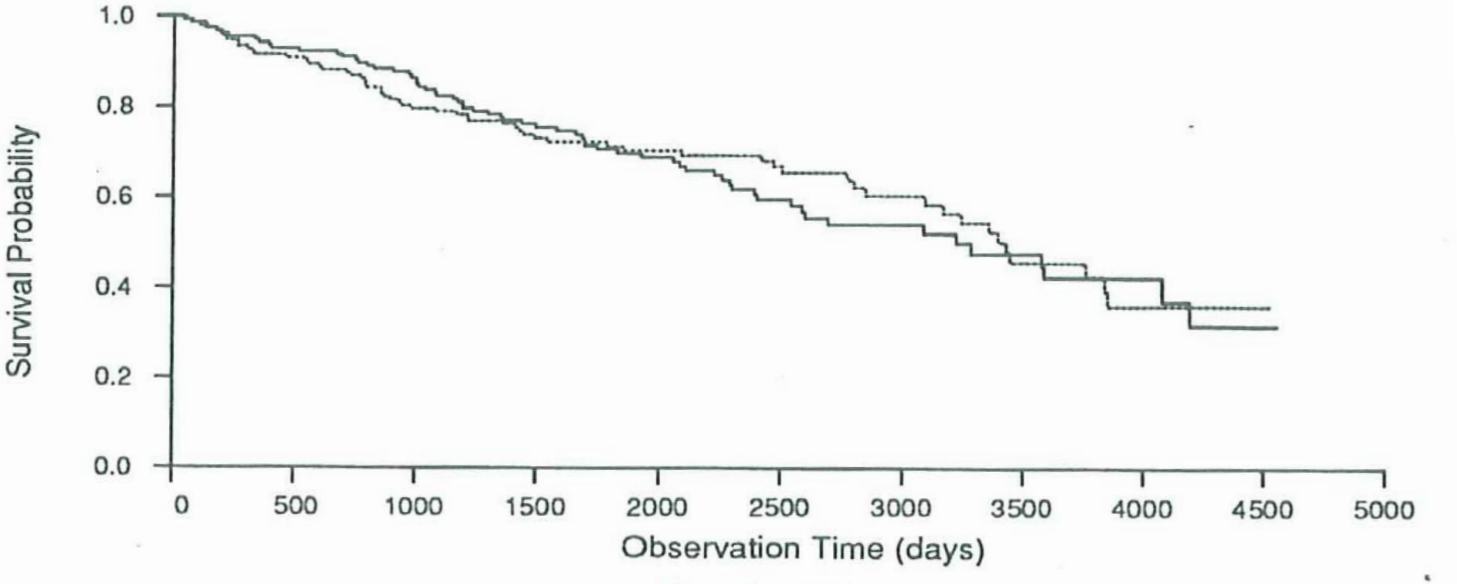
Log OR = -0.511 ± 0.305 unadjusted

Log OR = -0.616 ± 0.335 adjusted for maternal age, no. previous miscarriages, previous live birth

# Mesothelioma Trial (ASCO, 2006)



# MAYO Clinic PBC Trial (Fleming and Harrington, Wiley, 1991)




| Group | Time Interval 0-1000 | 1000-2000 | 2000-3000 | 3000-4000 | 4000-5000 |
|---|---|---|---|---|---|
| DPCA | 23/158 | 22/128 | 13/74 | 5/31 | 2/10 |
| Placebo | 31/154 | 12/120 | 7/70 | 10/32 | 0/11 |

(# events/# at risk)

# MAYO Clinic PBC Trial (cont.)

Unadjusted: Log HR = -0.0571 ± 0.1792

**Table 4.4.4 Adjusted estimation of treatment effect, 312 randomized cases with PBC.**

(a) Log likelihood −540.144

| | Coef. | Std. Err. | Z stat. |
|---|---|---|---|
| Age | 0.0347 | 0.00891 | 3.89 |
| log(Albumin) | −3.0771 | 0.71899 | −4.28 |
| log(Bilirubin) | 0.8840 | 0.09871 | 8.96 |
| Edema | 0.7859 | 0.29647 | 2.65 |
| log(Prothrombin Time) | 2.9707 | 1.01588 | 2.92 |
| Treatment | 0.1360 | 0.18543 | 0.73 |

## Logistic Regression Model

$$\log\left(\boldsymbol{p}/(1-\boldsymbol{p})\right) = \boldsymbol{a}^{*} + \boldsymbol{b}_1^{*}\boldsymbol{Z}_1 ,$$

$\boldsymbol{Z}_1 =$ treatment

$\boldsymbol{X}_2 =$ covariate

$$\log\left(\boldsymbol{p}/(1-\boldsymbol{p})\right) = \boldsymbol{a} + \boldsymbol{b}_1\boldsymbol{Z}_1 + \boldsymbol{b}_2\boldsymbol{X}_2$$

Robinson and Jewell (Int. Stat. Review, 1991): $\boldsymbol{Var}(\hat{\boldsymbol{b}}_1) \geq \boldsymbol{Var}(\hat{\boldsymbol{b}}_1^{*})$

(no improvement in precision)

Cox Regression Model?

“ The introduction of covariates may have negligible effect on the estimated standard errors of estimated treatment effects. In fact, it may increase them.”

Ford, Norrie, Ahmadi (SIM, 1995)

# UC Randomized Clinical Trial of Nitric Oxide in Premature Infants (NOVA)

Schreiber et al. (New Eng J Med, 2003)

Nitric oxide attenuates pulmonary vascular disease, inflammation, and pulmonary hypertension in newborns with lung injury.

Hypothesis: Use of inhaled nitric oxide would decrease the incidence of chronic lung disease and death in premature infants with respiratory distress syndrome (RDS).

Patient population: Infants < 72 hrs old with RDS
< 34 wks gestation
< 2000 g
No major congenital malformations

# NOVA trial (cont.)

2x2 factorial design:

| | INO | Placebo |
|---|---|---|
| Conventional Ventilation | n=51 | n=54 |
| Hi-freq. Oscillatory Ventilation | n=54 | n=48 |
| | 105 | 102 |

Randomization was stratified by birth weight.

# NOVA trial (cont.)

| | INO (n=105) | Placebo (n=102) |
|---|---|---|
| Death or CLD | 51 (48.6%) | 65 (63.7%) |
| Survived w/o CLD | 54 (51.4%) | 37 (36.3%) |
| | 105 | 102 |

# Logistic Regression Analysis

| Unadjusted | Coef. | Std. Err. | z | P>\|z\| |
|---|---|---|---|---|
| Treatment | -.6206 | .2838 | 2.19 | 0.029 |

| Adjusted | Coef. | Std. Err. | z | P>\|z\| |
|---|---|---|---|---|
| Treatment | -.6620 | .3422 | 1.93 | 0.053 |
| Sex | -1.1250 | .3752 | -3.00 | 0.003 |
| Birthwt* | -.2677 | .0865 | -3.10 | 0.002 |
| Gestage | -.1820 | .1128 | 1.61 | 0.107 |
| Race | -.2460 | .3755 | -0.65 | 0.512 |
| HiFreq | .2122 | .3411 | 0.62 | 0.534 |

* Per 100 gm

# Logistic Regression Analysis (cont.)

Unadjusted SE = 0.2838

| Adjusted | Coef. | Std. Err. | z | P>\|z\| |
|---|---|---|---|---|
| Treatment | -.6290 | .3365 | 1.87 | 0.062 |
| Sex | -1.2174 | .3638 | -3.35 | 0.001 |
| Birthwt* | -.3712 | .0596 | -6.23 | 0.000 |

* Per 100 mg

| Adjusted | Coef. | Std. Err. | z | P>\|z\| |
|---|---|---|---|---|
| Treatment | -.6900 | .3295 | 2.09 | 0.036 |
| Sex | -.7864 | .3393 | -2.32 | 0.020 |
| Gestage | -.4579 | .0748 | -6.12 | 0.000 |

**Time, effort, $**

# Solutions

Data dredging problem:

- Stipulate exactly what you are going to do in the protocol
- Tukey's (CCT, 1993) 'compound covariates' method
- Propensity scores
- Tsiatis, Davidian et al. (SIM, 2007) 'semi-parametric theory'
  "decouples evaluation of the treatment effect from regression modeling"

Lack of improvement in precision for non-linear models:

- Tsiatis, Davidian et al. approach

## Propensity Score?

Usually used to reduce bias in observational studies

-- Rosenbaum and Rubin (JASA, 1984)

Propensity score for i$^{th}$ subject is the conditional probability of assignment to a particular treatment ($Z_i = 1$) versus control ($Z_i = 0$) given a vector of covariates $X_i$.

$$e(x_i) = pr(Z_i = 1 \mid X_i = x_i)$$

No reason why this can't be used in RCTs (Abramowski, SCT 1991).

Outcome variable, $Y_i$ , does not enter into modeling process.

# NOVA trial

Logistic regression of *treatment assignment* on covariates:

| Parameter | Estimate | Standard Error | z | Pr > \|z\| |
|---|---|---|---|---|
| SEX | -0.1240 | 0.2966 | 0.03 | 0.973 |
| BIRTHWT* | 0.0474 | 0.0690 | 0.69 | 0.675 |
| Gestage | -0.0045 | 0.0955 | 0.05 | 0.962 |
| racec | -0.2099 | 0.3078 | 0.68 | 0.495 |
| HIFREQ | 0.1606 | 0.2845 | 0.56 | 0.572 |

* Per 100 gm

$\Rightarrow \hat{e}(x_i)$ propensity score (PS)

Unadjusted SE = 0.2838

Treatment effect adjusted for propensity score:

| Parameter | Estimate | Standard Error | z | Pr > \|z\| |
|---|---|---|---|---|
| Treatment | -0.5331 | 0.2955 | 1.80 | 0.071 |
| PS* | 1.0317 | 0.2695 | 3.83 | 0.0001 |

* Per 0.1 increase

| Parameter | Estimate | Standard Error | z | Pr > \|z\| |
|---|---|---|---|---|
| Treatment | -0.5463 | 0.2966 | 1.84 | 0.066 |
| quintile 1 | -0.8776 | 0.2970 | 2.95 | 0.003 |
| quintile 2 | -0.2963 | 0.2838 | 1.04 | 0.296 |
| quintile 3 | 0.2314 | 0.2918 | 0.79 | 0.428 |
| quintile 4 | 0.1229 | 0.2867 | 0.43 | 0.668 |

Mantel-Haenszel: log OR = -0.5242, SE = 0.3037 p=0.084

Adjustment based on the propensity score also “decouples evaluation of the treatment effect from regression modeling.”

Focuses on adjusting for chance imbalances in covariates between treatment arms (in RCT), not necessarily covariates that are predictive of outcome.

-- Consequently, less likely to improve precision.

Standard logistic regression analysis *will not* increase precision of estimated log OR. Robinson and Jewell (1991)

## Tsiatis, Davidian et. al approach

Makes use of semi-parametric theory (Tsiatis, Springer 2006)

Unconditional treatment effect:

$$\beta = \boldsymbol{E}(\boldsymbol{Y} \mid \boldsymbol{Z} = 1) - \boldsymbol{E}(\boldsymbol{Y} \mid \boldsymbol{Z} = 0)$$

Key point: this is a difference in *means.* (For binary outcomes, difference in proportions rather than log OR.)

Also, in non-linear models, $\beta$ is generally different from the conditional treatment effect:

$$\beta_{\boldsymbol{x}} = \boldsymbol{E}(\boldsymbol{Y} \mid \boldsymbol{Z} = 1, \boldsymbol{X} = \boldsymbol{x}) - \boldsymbol{E}(\boldsymbol{Y} \mid \boldsymbol{Z} = 0, \boldsymbol{X} = \boldsymbol{x})$$

All “reasonable and asymptotically normal estimators for $\beta$ “ can be expressed as

$$\bar{Y}^{(1)} - \bar{Y}^{(0)} - \sum_{i=1}^{n} (Z_i - \bar{Z}) \left\{ n_0^{-1} h^{(0)}(X_i) + n_1^{-1} h^{(1)}(X_i) \right\}, \quad (1)$$

where $n = n_0 + n_1$ and $h^{(k)}(X), k = 0, 1,$ are arbitrary scalar functions of $X$.

Note: ANCOVA model is a special case with

$$h^{(0)}(X_i) = h^{(1)}(X_i) = \Sigma_{XY}^{T} \Sigma_{XX}^{-1} X_i$$

Main result: Among *all* estimators exactly equal to or asymptotically equivalent to (1), that with the smallest variance is

$$\boldsymbol{h}^{(\boldsymbol{k})}(\boldsymbol{X}_{\boldsymbol{i}}) = \boldsymbol{E}(\boldsymbol{Y}_{\boldsymbol{i}} \mid \boldsymbol{Z}_{\boldsymbol{i}} = \boldsymbol{k}, \boldsymbol{X}_{\boldsymbol{i}}), \boldsymbol{k} = 0,1$$

The "optimal" functions ***h*** are the *true* regression relationships of Y on X for each treatment separately.

These may be non-linear in X and different for the two treatments.

Moreover, if one restricts ***h*** to linear models, even if the true relationship is not linear, the variance of the resulting estimator for $\beta$ will be *less than* the variance of the unadjusted estimator.

Note: rather than regressing Y on both Z and X,

$\boldsymbol{E}(\boldsymbol{Y} \mid \boldsymbol{Z} = \boldsymbol{k}, \boldsymbol{X}), \boldsymbol{k} = 0,1$ are fitted separately by treatment.

Decouples evaluation of the treatment effect from regression modeling.

Four-step algorithm:

(1) Partition the data into two sets by treatment group.

(2) For each set, model $\boldsymbol{E}(\boldsymbol{Y} \mid \boldsymbol{Z} = \boldsymbol{k}, \boldsymbol{X})$. Achieve as good a fit as possible: can include squared terms and interactions among the covariates.

(3) Denote the models by $\boldsymbol{f_k}(\boldsymbol{X}, \alpha_{\boldsymbol{k}}), \boldsymbol{k} = 0,1$ and for each

$$i = 1,2,\ldots,n \text{ form } \hat{f}_{0,i} = f_0(X_i, \hat{\alpha}_0) \text{ and } \hat{f}_{1,i} = f_1(X_i, \hat{\alpha}_1).$$

(4) Calculate

$$\hat{\beta} = \overline{Y}^{(1)} - \overline{Y}^{(0)} - \sum_{i=1}^{n} (Z_i - \overline{Z})(n_0^{-1}\hat{f}_{0,i} + n_1^{-1}\hat{f}_{1,i})$$

and

$$Var(\hat{\beta}) = C\sum_{i=1}^{n}\Big[\Big\{n_1^{-1}Z_i - n_0^{-1}(1-Z_i)\Big\}Y_i - n^{-1}\hat{\beta} - (Z_i - \overline{Z})(n_0^{-1}\hat{f}_{0,i} + n_1^{-1}\hat{f}_{1,i})$$

$$-(Z_i - \overline{Z})\Big\{n_0^{-1}(\overline{Y}^{(0)} - \bar{f}_0) + n_1^{-1}(\overline{Y}^{(1)} - \bar{f}_1)\Big\}\Big]^2$$

***C*** is a correction factor and $\bar{f}_k, k = 0,1$ are the means of the predicted values for each group.

# NOVA Trial

| | INO (n=105) | Placebo (n=102) |
|---|---|---|
| Death or CLD | 51 (48.6%) | 65 (63.7%) |
| Survived w/o CLD | 54 (51.4%) | 37 (36.3%) |
| | 105 | 102 |

Unadjusted treatment effect:

$$\bar{Y}^{(1)} - \bar{Y}^{(0)} = 51/105 - 65/102 = .4857 - .6373 = -.1515.$$

$$SE(\bar{Y}^{(1)} - \bar{Y}^{(0)}) = \sqrt{\frac{.4857(1-.4857)}{105} + \frac{.6373(1-.6373)}{102}} = 0.0682$$

## Adjusted Estimate

In treated group, fit logistic regression model with covariates:

sex, birthweight, gestage, race, and HiFreq

Used same model for control group.

Note:

$$\hat{\boldsymbol{f}}_{0,\boldsymbol{i}} = \frac{\exp(\hat{\alpha}_0^{\boldsymbol{T}} \boldsymbol{X}_{\boldsymbol{i}})}{1+\exp(\hat{\alpha}_0^{\boldsymbol{T}} \boldsymbol{X}_{\boldsymbol{i}})},\ \boldsymbol{i} = 1,2,\ldots,\boldsymbol{n}$$

$$\hat{\boldsymbol{f}}_{1,\boldsymbol{i}} = \frac{\exp(\hat{\alpha}_1^{\boldsymbol{T}} \boldsymbol{X}_{\boldsymbol{i}})}{1+\exp(\hat{\alpha}_1^{\boldsymbol{T}} \boldsymbol{X}_{\boldsymbol{i}})},\ \boldsymbol{i} = 1,2,\ldots,\boldsymbol{n}$$

Model fits?

Hosmer-Lemeshow statistics

Treated arm: $\chi^2$ = 9.79 on 8 df

Control arm: $\chi^2$ = 3.98 on 8 df

Result: $\hat{\beta} = -0.1158$

$SE(\hat{\beta}) = 0.0590$

which compares favorably with unadjusted SE = 0.0682

(although p-value increases from 0.026 to 0.050).

# Conclusions

- It is important to present both unadjusted and adjusted comparisons.
- The unadjusted comparison in a RCT is defensible and avoids any concerns about *post hoc* fishing for the “most satisfactory” p-value.
- However we will have to deal with critics who spot imbalances, so some adjustment seems inevitable.
- Also want to improve efficiency. Having spent valuable resources on collecting covariate information, we should take advantage of it.

## Are adjustments worth the added complexity?

Until now, depends on the model . . .

Linear models: Yes, precision is improved.

Non-linear models: Less clear, precision may not be improved, although may still prefer conditional estimate.

Legitimate concerns about the analyst selecting the most favored result.

Method developed by Tsiatis, Davidian et al. offers a way to avoid the data dredging problem AND increase precision for both linear and non-linear models

NOVA: $$\boldsymbol{EFF} = \frac{.0682^2}{.0590^2} = 1.34$$

# Further work:

For binary outcomes, can we improve upon estimates of the (log) odds ratio by incorporating covariates?

Zhang, Tsiatis, and Davidian (Biometrics, 2008)

Hazard ratio for time-to-event data?

Lu and Tsiatis (Biometrika, 2008)

Question: What about covariates that are imbalanced between the two treatment arms, but only mildly correlated with outcome? These may not be picked up by the independent analysts. The adjusted result may not therefore satisfy critics who raise concerns about such imbalances.

# Thank you!