

The **BIG** Biostat Picture

Intro Biostatistics: *a bit of everything...*

- Data summaries (means, medians...)
- EDA (Exploratory data analysis)
- CDA (Confirmatory data analysis): 1-sample inference
 - b hypothesis testing
 - ▷ significance
 - ⊳ power
- CDA: 2-sample inference
 - > means
 - proportions
- Intro to linear regression

The **BIG** Biostat Picture

Intro to Regression: *continuous response variables*

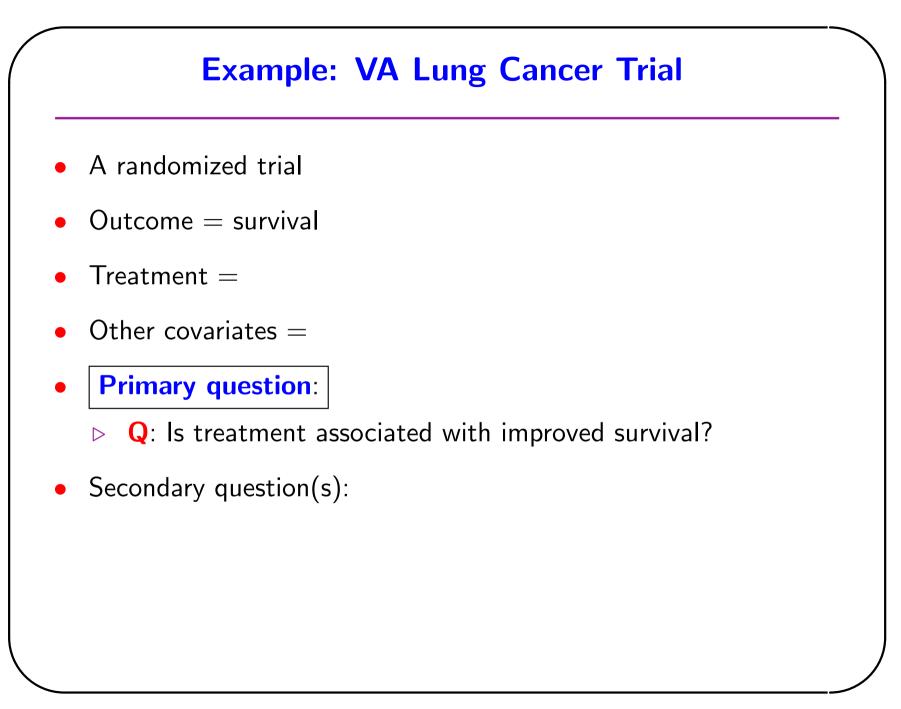
- Simple linear regression
- Transformation(s) both Y and X
- Residuals
- Multiple regression
- Confounding & Interaction
- Diagnostics
- Dummy variables
- ANOVA, ANCOVA
- Repeated measures

The **BIG** Biostat Picture

This course: *survival outcomes*

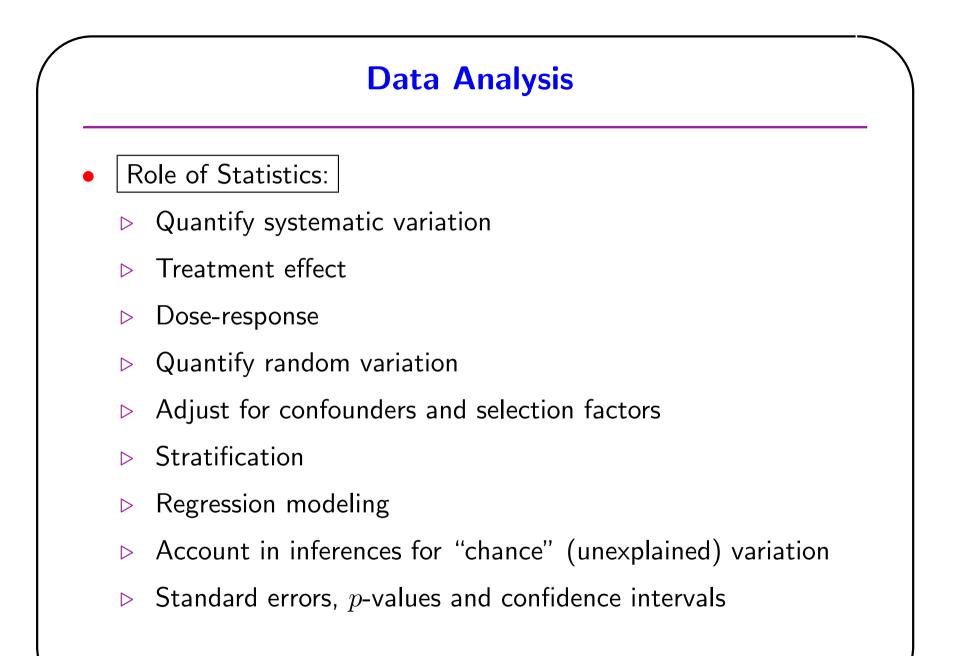
- Survival Data censoring
- Estimation of Survival Curves (Kaplan-Meier)
- Comparing Survival Curves (log rank test)
- Regression for Survival Data: Cox Regression
 - \triangleright Single binary predictor, X
 - \triangleright Multiple predictors, X_1, X_2, \ldots
 - Regression for adjustment
 - ▷ Regression for prediction

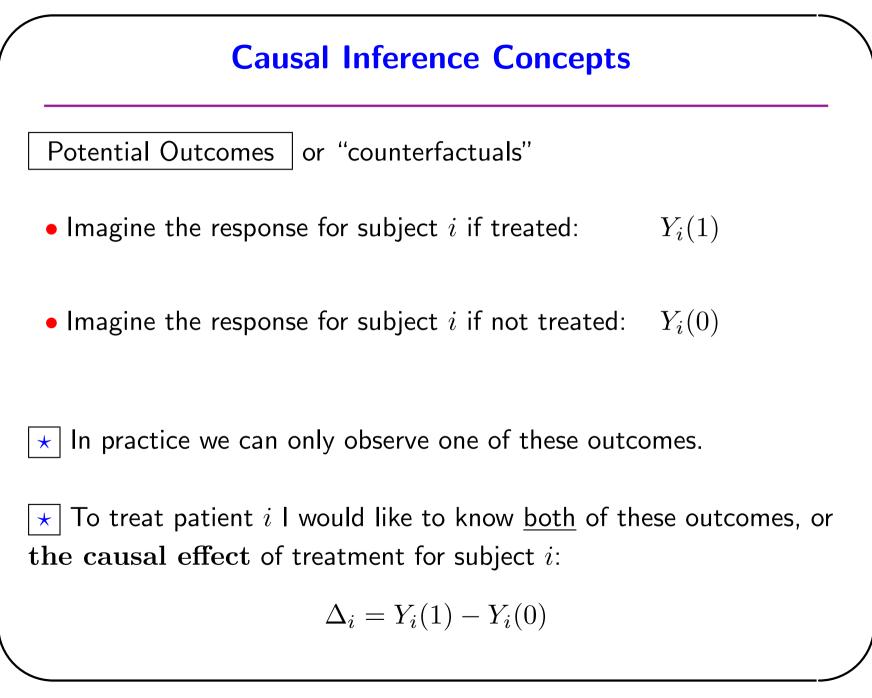




Data Analysis

"In an analysis, the basic questions to consider are the degree of association between risk for disease and the factors under study, the extent to which the observed associations may result from bias, **confounding**, and/or **chance**, and the extent to which they may be described as **causal**." (*Breslow and Day (1980) Vol. I*)





• In a randomized study we can get an unbiased estimate of the **average causal effect**.

	Treatment	Potential Outcomes	
Subject	Assignment	Tx=1	Tx=0
1	0		$Y_{1}(0)$
2	1	$Y_{2}(1)$	
3	1	$Y_3(1)$	
4	0		$Y_4(0)$
5	1	$Y_5(1)$	
6	0		$Y_6(0)$
7	0		$Y_7(0)$
8	1	$Y_8(1)$	
		\overline{Y}_1	\overline{Y}_0

• Outcome = Pain Scale, 0-10 points (high is worse)

	Baseline	Potential Outcomes		
Subject	Status	Tx=1	Tx=0	Δ
1	S	7	9	-2
2	S	6	9	-3
3	S	5	8	-3
4	S	3	7	-4
5	NS	3	6	-3
6	NS	2	5	-3
7	NS	3	4	-1
8	NS	2	4	-2
		$\mu(1) = 3.9$	$\mu(0) = 6.5$	$\overline{\Delta} = -2.6$

• Conclusion?

Estimation

(1) Because we randomize \overline{Y}_1 is an unbiased estimate of the average response for the population if everyone was treated, $\mu(1)$. That is,

$$E\left[\overline{Y}_1\right] = \mu(1) = \frac{1}{N} \sum_{i=1}^N Y_i(1)$$

(2) Because we randomize $\overline{Y}(0)$ is an unbiased estimate of the average response for the population if everyone was not treated, $\mu(0)$. That is,

$$E\left[\overline{Y}_0\right] = \mu(0) = \frac{1}{N} \sum_{i=1}^N Y_i(0)$$

(3) This implies that we can estimate the average causal effect:

$$\widehat{\Delta} = \overline{Y}_1 - \overline{Y}_0$$

$$E\left[\widehat{\Delta}\right] = E\left[\overline{Y}_{1}\right] - E\left[\overline{Y}_{0}\right]$$

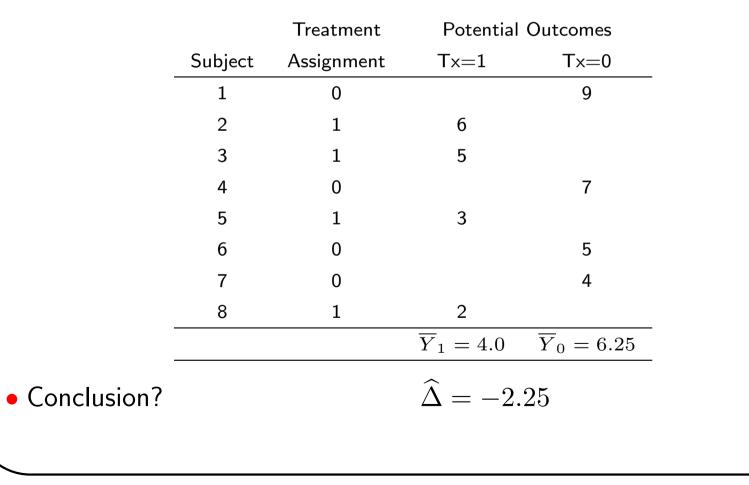
$$= \frac{1}{N}\sum_{i=1}^{N}Y_{i}(1) - \frac{1}{N}\sum_{i=1}^{N}Y_{i}(0)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left[Y_{i}(1) - Y_{i}(0)\right] = \frac{1}{N}\sum_{i=1}^{N}\Delta_{i}$$

$$= \text{ average causal effect }, \overline{\Delta}$$

• Outcome = Pain Scale, 0-10 points (high is worse)

• Treatment = Surgery (Tx=1) or Conservative (Tx=0)

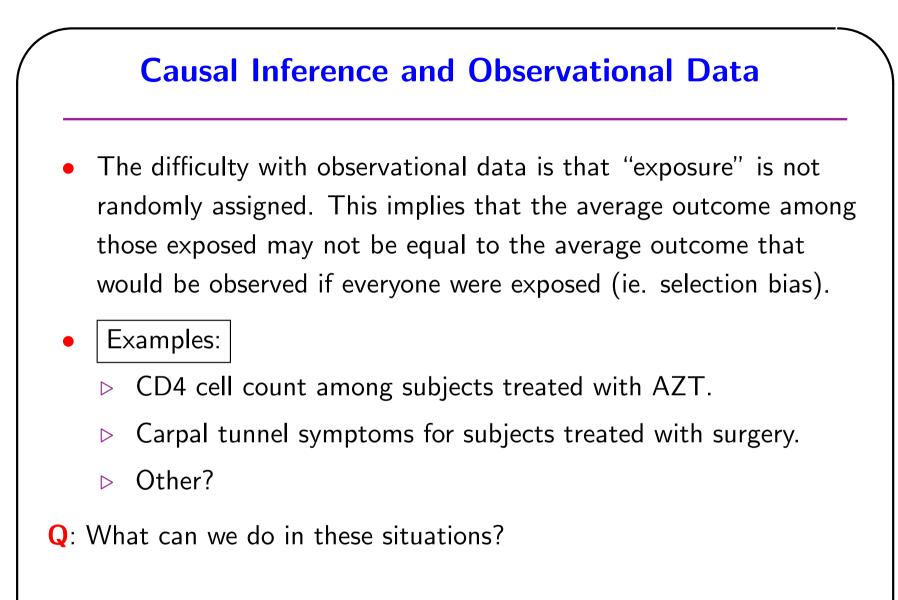


Estimation

- We can estimate average causal effects when there is nothing (else) that systematically differs between the exposed and the unexposed.
- Randomization guarantees "no unmeasured confounding".
- There is no single "effect" of exposure since for each individual we have a possibly different exposure effect, Δ_i .
- Different populations (ie. young age, old age) may have different average causal effects.

Observational Studies

Q: Can we estimate causal effects based on observational data?



A: Control for factors via stratification or regression adjustment.

P. Heagerty, VA/UW Summer 2005

Observed Data

	Baseline	Potential Outcomes		
Subject	Status	Tx=1	Tx=0	Tx Received
1	S	7		1
2	S	6		1
3	S	5		1
4	S		7	0
5	NS		6	0
6	NS		5	0
7	NS	3		1
8	NS		4	0
		$\overline{Y}_1 = 5.25$	$\overline{Y}_0 = 5.5$	$\widehat{\Delta} = -0.25$

• Conclusion?

Stratify on Baseline:

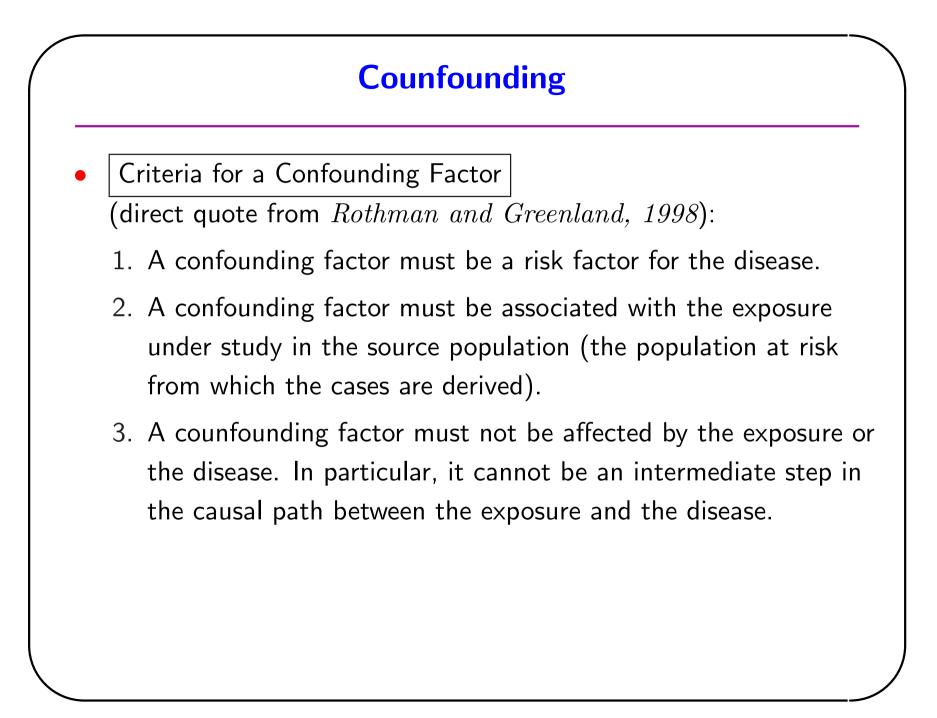
	Baseline	Potential	Outcomes	
Subject	Status	Tx=1	Tx=0	Tx Received
 1	S	7		1
2	S	6		1
3	S	5		1
4	S		7	0
		$\overline{Y}_1 = 6.0$	$\overline{Y}_0 = 7.0$	$\widehat{\Delta}_S = -1.0$
 5	NS		6	0
6	NS		5	0
7	NS	3		1
8	NS		4	0
		$\overline{Y}_1 = 3.0$	$\overline{Y}_0 = 5.0$	$\widehat{\Delta}_{NS} = -2.0$
				$\widehat{\Delta}=$ -1.5

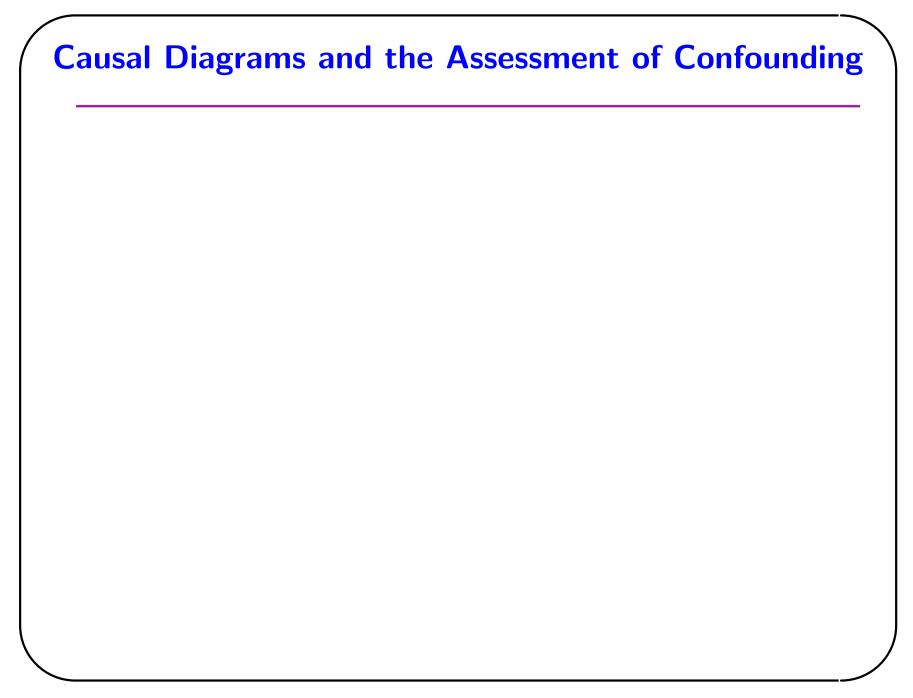
Confounding

"A confounding variable is a variable that is associated with both the disease and the exposure variable." Rosner (1995)

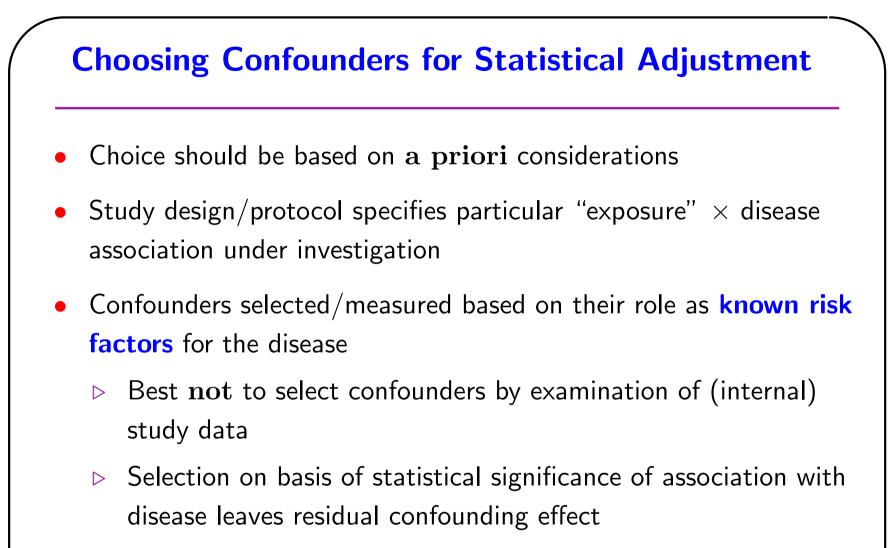
"Confounding is the distortion of a disease/exposure association brought about by the association of other factors with both disease and exposure, the latter associations with disease being causal." Breslow & Day (1980)

"If any factor either increasing or decreasing the risk of a disease besides the characteristic or exposure under study is unequally distributed in the groups that are being compared with regard to the disease, this itself will give rise to differences in disease frequency in the compared groups. Such distortion, termed confounding, leads to an invalid comparison." *Lilienfeld & Stolley (1994)*

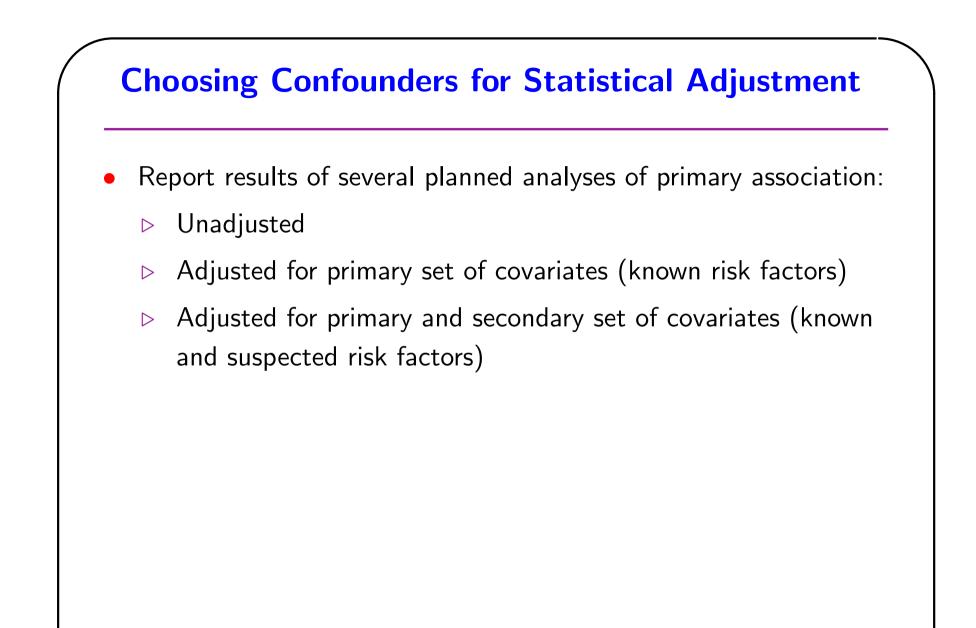




P. Heagerty, VA/UW Summer 2005



Selection on basis of resulting change in exposure/disease effect measure destroys opportunity for correct inferences



Example: VA Lung Cancer Trial					
Q : Can we use <u>linear regression</u> to compa groups?	re treatment and control				
Outcome:					
Model(s):					

	Example: VA Lung Cancer Trial				
g	Q : Can we use <u>logistic regression</u> to compare treatment and control roups?				
С	Outcome:				
N	lodel(s):				

Summary

- Regression methods allow inference regarding "exposures" for a variety of outcomes.
- Regression can be used to adjust for confounding variables.
- There is no single effect of the exposure only average effects.
- Linear regression continuous outcome; means; differences in means.
- Logistic regression binary outcome; log odds; log odds ratios.
- 🖈 Cox regression censored survival; hazard; hazard ratios.