

# Confounding, interaction, and mediation in multivariable/multivariate regression modeling

William Wu

Department of Biostatistics  
Cancer Biostatistics Center, Vanderbilt-Ingram Cancer Center

May 21, 2010

# Outline

## 1 Mediation

3 examples

Definition and identification

An application

## 2 Confounding

Definition and determination

Methods for confounding effect

Unobserved confounding

Difference from mediation

## 3 Interaction

Definition

Determination of interaction

Difference from mediation and confounding

# MEDIATION

## Pingsheng's study

### Study question:

Whether and how childhood asthma was associated with maternal smoking and infancy bronchiolitis.

### Preliminary modeling finding:

The significant association between maternal smoking and asthma was found, but the association was gone after adjusting for bronchiolitis in multivariable modeling.

*"We hypothesize that one mechanism through which maternal smoking during pregnancy contributes to the known increased risk of developing childhood asthma is through increasing the risk of an important intermediate event, bronchiolitis during infancy."*

## Dan Weeks' talk

In the paper:

- *'Interpretation of Genetic Association Studies: Markers with Replicated Highly Significant Odds Ratios May Be Poor Classifiers.'*
- *'Although a set of SNPs can be strongly associated with disease risk with extremely small P-values, the same set of SNPs may not necessarily have high discrimination ability or may not dramatically improve the discrimination ability of a classification model constructed using 'conventional' non-genetic risk factors without the SNPs.'*

*PLoS genetics 2009;5(2)*

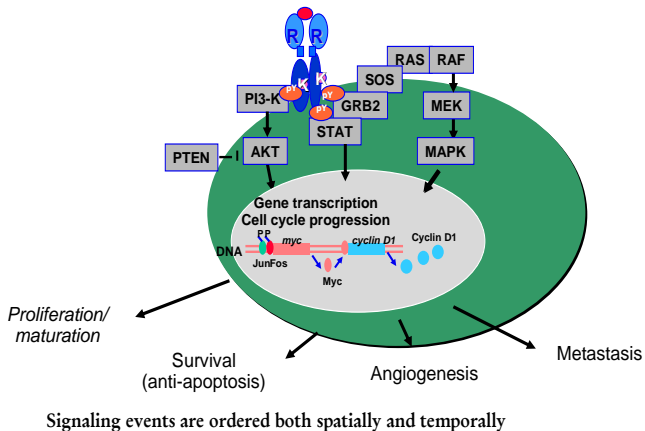
# Adriana Gonzales' study

Study question:

How biomarkers EGFR, AKT, and Ki-67 were correlated among 69 osteosarcoma patients.

*H Wu et al. Biomarker Insights 2007;2:469-76*

## EGFR pathway



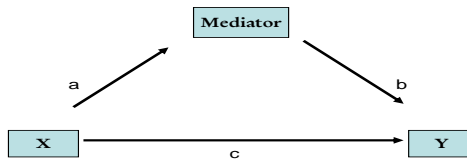
## Possible approaches to Adriana's data

- Correlation analysis of the 3 biomarkers?
- Regression of Ki-67 on other 2 biomarkers?

# What is mediation?

A mediation effect occurs when the third variable (mediator, M) carries the influence of a given independent variable (X) to a given dependent variable (Y).

Mediation models explain how an effect occurred by hypothesizing a causal sequence.

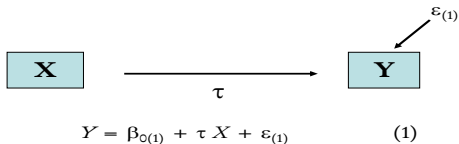


## Approaches to identification

- Approach 1: Causal steps
- Approach 2: Statistical test

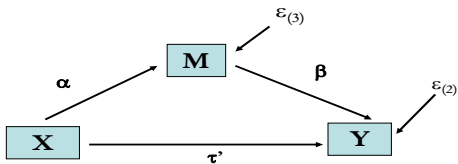
# Approach 1: causal steps

## Model 1



# Causal steps

## Model 2 and Model 3



$$Y = \beta_{0(2)} + \tau' X + \beta M + \varepsilon_{(2)} \quad (2)$$

$$M = \beta_{0(3)} + \alpha X + \varepsilon_{(3)} \quad (3)$$

## A significant mediation effect should be:

- $\tau$  in Model 1  
The total effect of the independent variable X on the dependent variable Y must be significant.
- $\alpha$  in Model 3  
The path from X to M must be significant.
- $\beta$  in Model 2  
The path from M to Y must be significant.
- $\tau'$  in Model 2  
Evidence for mediation when  $\tau'$  becomes insignificant when the M is included (effect of X on Y is zero). This would be complete mediation

## Approach 2: statistical test of mediation

Sobel test:

to test the products of coefficients of the two paths  $a$  and  $b$ .

$$z - \text{value} = \alpha * \beta / \sqrt{\alpha^2 \sigma_\beta^2 + \beta^2 \sigma_\alpha^2}$$

The null hypothesis is a test of  $\alpha * \beta = 0$ .

*MacKinnon and Dwyer (1994) and MacKinnon, Warsi, and Dwyer (1995)*

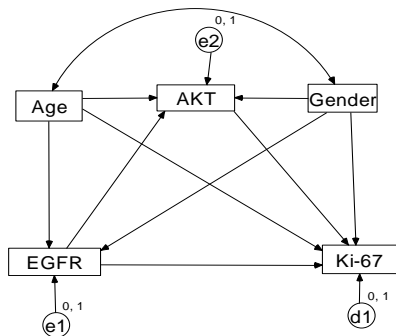
## Adriana's study

- Initial variable EGFR and mediating variable Akt were immunostaining index.
- Outcome variable Ki-67 was a cancer cell proliferation index and also an immunostaining index.

## Approach 1: the 3 models

- $Model1 \leftarrow ols(Ki67 \sim EGFR + age + sex)$
- $Model2 \leftarrow ols(Ki67 \sim EGFR + AKT + age + sex)$
- $Model3 \leftarrow ols(AKT \sim EGFR + age + sex)$

# Joint modeling (SEM)



## Approach 1: results

| Model                 | Coefficient          | SE        | F     | p      |
|-----------------------|----------------------|-----------|-------|--------|
| ols (Ki67~ EGFR+)     | $\tau$ : 0.0003069   | 0.0001400 | 4.80  | 0.0340 |
| ols (AKT~ EGFR+)      | $\alpha$ : 0.6584    | 0.1996    | 10.89 | 0.0020 |
| ols (Ki67~ EGFR+AKT+) | $\beta$ : 0.0003019  | 0.0000993 | 9.24  | 0.0042 |
| ols (Ki67~ EGFR+AKT+) | $\tau'$ : 0.00009687 | 0.0001444 | 0.45  | 0.5061 |

# Components of mediation model

- Total effect=

$$\alpha * \beta + \tau' = 0.6584 \times 0.0003019 + 0.00009687 = 0.0002957 (\cong \tau = 0.0003069)$$

- Direct effect=

$$\tau' = 0.00009687$$

- Mediated effect=

$$\alpha * \beta = 0.6584 \times 0.0003019 = 0.0001988 =$$

$$(\text{total} - \text{direct} = \tau - \tau' = 0.0002957 - 0.00009687 = 0.0001988)$$

## Approach 2: results

| Test         | <i>p</i> value |
|--------------|----------------|
| Sobel        | 0.00000152     |
| Goodman (I)  | 0.00000172     |
| Goodman (II) | 0.00000133     |

***A significant mediating effect for Akt was found with the tests.***

# CONFOUNDING

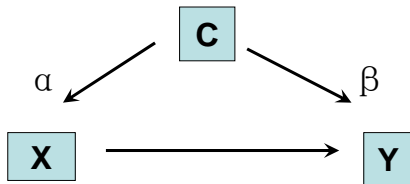
# What is a confounder?

## Criteria for a confounder

- 1 It is a risk factor for the disease, independent of the putative risk factor (exposure variable or X).
- 2 It is associated with putative risk factor (exposure).
- 3 It is not in the causal pathway between exposure and disease.

## Confounding model

The association between  $X$  (exposure) and  $Y$  (outcome) is distorted by the presence of another variable  $C$  (confounder)



## An example

- Age may confound the positive relationship between annual income and cancer incidence in the US.
  - ① Older individuals are also more likely to get cancer.
  - ② Older individuals are likely to earn more money than younger ones who have not spent as much time in the work force.
  - ③ Income does not cause age, which then causes cancer.

# Confounding

- Not an issue in randomized study  
Randomization process will eliminate the correlation between confounder and exposure, e.g., age and annual income (i.e. we should have roughly equal numbers of age category in each annual income group).
- An issue in observational study.

## Consequence of confounding

- Bias effect estimate.
- Widen confidence interval.
- Inclusion of additional potential confounders may only widen the CI and have no impact of effect estimate.
- Confounding is the masking of the true effect of a risk factor on a disease or outcome by the presence of another variable.
- The presence of confounding effect leads to a spurious association between exposure and outcome.

# How to pick potential confounders?

## philosophically

- Not a simple question
- Your knowledge
- Prior experience with data
- The three criteria for confounders

# How to pick potential confounders?

statistically

- When you get to doing multivariable logistic regression, for example, one rule of thumb is that if the odds ratio changes by 10% or more then this is reason to include the potential confounder in your multi-variable model. We don't tend to look at just whether it is statistically significant, but instead, how much does it change with this effect. This change is what we want to measure. If it changes the effect by 10% or more, then we consider it a confounder and leave it in the model.
- $P$  values will not tell confounding effect. Rather, change in  $\beta$  between with adjustment and w/o adjustment can tell if the confounding is working.

## Design stage

- ① Randomization
- ② Restriction (of the study population to a category of a confounder, but will limit generalizability)
- ③ Matching (Not feasible matching all, residual confounding will still bias estimate)

## Data analysis stage

- 1 Adjustment (a few to dozen)
- 2 propensity score (all)

# What is Unobserved confounding?

Confounding that remains after adjustment for observed confounders are referred as residual confounding. This involves unmeasured confounding as well as inaccurately measured confounding.

## Methods for unobserved confounding

- A challenge  
can not adjust for  
only randomization
- But you can  
adjust for as many as allowed  
improve the measurement (Measurement error in confounders  
will lead to residual confounding).  
use more appropriate scaling of the measurement  
more ...

# Other proposed approaches for unobserved confounding

BIOMETRICS 56, 915-921 September 2000

## When Should Epidemiologic Regressions Use Random Coefficients?

Sander Greenland

Department of Epidemiology, UCLA School of Public Health,  
Los Angeles, California 90095-1772, U.S.A.

**SUMMARY.** Regression models with random coefficients arise naturally in both frequentist and Bayesian approaches to estimation problems. They are becoming widely available in standard computer packages under the headings of generalized linear mixed models, hierarchical models, and multilevel models. I here argue that such models offer a more scientifically defensible framework for epidemiologic analysis than the fixed-effects models now prevalent in epidemiology. The argument invokes an antiparsimony principle attributed to L. J. Savage, which is that models should be rich enough to reflect the complexity of the relations under study. It also invokes the countervailing principle that you cannot estimate anything if you try to estimate everything (often used to justify parsimony). Regression with random coefficients offers a rational compromise between these principles as well as an alternative to analyses based on standard variable-selection algorithms and their attendant distortion of uncertainty assessments. These points are illustrated with an analysis of data on diet, nutrition, and breast cancer.

## Confounding is different from mediation in:

- Temporality (Exposure occurs first and then M and outcome, and conceptually follows an experimental design)
- Directionality
- Causality
- Confounders often demographic variables that typically cannot be changed in an experimental design. Mediators are by definition cable of being changed and are often selected based on malleability.
- statistical test

# statistical test for confounding

TECHNICAL REPORT  
R-256  
January 1998

## WHY THERE IS NO STATISTICAL TEST FOR CONFOUNDING, WHY MANY THINK THERE IS, AND WHY THEY ARE ALMOST RIGHT\*

**Judea Pearl**Cognitive Systems Laboratory  
Computer Science Department  
University of California, Los Angeles, CA 90024  
*judea@cs.ucla.edu*

### 1 INTRODUCTION

Confounding is a simple concept. If we undertake to estimate the effect of one variable ( $X$ ) on another ( $Y$ ) by examining the statistical association between the two, we ought to ensure that the association is not produced by factors other than the effect under study. The presence of spurious association, due for example to the influence of extraneous variables, is called *confounding* as it tends to confound our reading and to bias our estimate of the effect studied. Conceptually, therefore, we can say that  $X$  and  $Y$  are confounded when there is a third variable  $Z$  that influences both  $X$  and  $Y$ ; such a variable is then called a “confounder” of  $X$  and  $Y$ .

As simple as this concept is, it has resisted formal treatment for several decades, and for a good reason: The very notions of “effect” and “influence”, relative to which “spurious association” is to be defined, has resisted mathematical formulation. The empirical definition of effect as an association that would prevail in a controlled randomized experiment, cannot easily be expressed in the standard language of probability theory, because that theory deals with static conditions, and does not permit us to predict, even from a full specification of a population density function, what relationships would prevail if conditions were to change, say from observational to controlled studies. Such predictions require extra information, in the form of causal or counterfactual assumptions [Greenland and Robins 1986; Wickramaratne and Holford 1987], which is not discernible from density functions.

These difficulties notwithstanding, epidemiologists, biostatisticians, social scientists and economists<sup>1</sup> have made numerous attempts to express confounding in statistical terms, partly

# INTERACTION

## What is interaction?

- An interaction means that the effect of X on Y depends on the level of a third variable.
- No causal sequence is implied by interaction.
- Also known as modification or moderation

## Understanding interaction effect

- We have the following regression on  $x_1$  and  $x_2$ :

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 * x_2) + \varepsilon$$

- The null hypothesis is  $H_0 : \beta_3 = 0$ , or product of the two variables,  $x_1$  and  $x_2$ , has no effect on  $Y$ .
- The test of  $H_0 : \beta_3 = 0$  is a test for parallelism of the two slopes (if  $x_2$  has two levels).

# Understanding interaction effect

- Given:

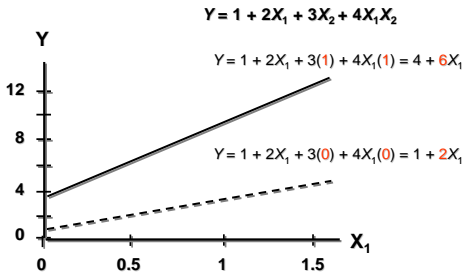
$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 * x_2) + \varepsilon$$

- Without interaction, effect  $x_1$  on  $y$  is measured by  $\beta_1$ .
- With interaction term, effect of  $x_1$  on  $y$  is measured by  $\beta_1 + \beta_3 x_2$ .

effect changes as  $x_2$  increases.

## When $x_2$ has two levels

Effect (slope) of  $X_1$  on  $Y$  does depend on  $X_2$  value.



## How to determine interaction?

philosophically

- Interaction can be an interest of the study
- Interaction is usually pre-specified

## An example about the philosophy

For example, imagine a study that tests the effects of a treatment of an outcome measure. The treatment variable is composed of two groups, e.g., treatment and control. The results are that the mean of the treatment group is higher than the mean for the control group. But what if the research is also interested in whether the treatment is equally effective for females and males. That is a difference in treatment depending on gender group. This is a question of interaction.

## How to determine interaction?

statistically

- Likelihood ratio test can be applied to test the interaction.
- Interaction terms can be excluded from the model if they are as a whole insignificant.
- Main effect may turn to be insignificant when interaction is included in model.
- Main effect won't tell the whole story in the presence of significant interaction.
- Stratified estimates are to be reported if the interaction is tested significant.

## A note

The statistical power to test the significant interaction is 5-times lower than that to test main effect. So,  $p = 0.10$  could be considered significant. Keep in mind we do not want to miss any important interaction.

## Interaction is different from mediation in:

- No causality
- Test for product of two measurements (test for product of two coefficients for mediation)
- Can be tested (confounding can not)

# Acknowledgements

Pingsheng Wu, Ph.D.  
Adriana Gonzalez, M.D., Ph.D.  
Debra Friedman, M.D., Ph.D.

Thank you!