# QUANTITATIVE BIAS ANALYSIS

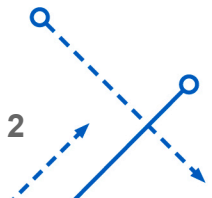
Hailey Banack, PhD

February 7, 2018

University at Buffalo The State University of New York

# Presentation Outline

- Motivation and key concepts
- What is quantitative bias analysis?
- Deterministic bias analysis:
  - Selection bias
  - Unmeasured confounders
  - Misclassification
- Multidimensional bias analysis
- Probabilistic bias analysis
- Advanced topics and extensions



# Motivating Example

ORIGINAL CONTRIBUTION

## Number of Coronary Heart Disease Risk Factors and Mortality in Patients With First Myocardial Infarction

John G. Canto, MD, MSPH
Catarina I. Kiefe, MD, PhD
William J. Rogers, MD
Eric D. Peterson, MD, MPH
Paul D. Frederick, MPH, MBA
William J. French, MD
C. Michael Gibson, MD
Charles V. Pollack Jr, MD, MA
Joseph P. Ornato, MD
Robert J. Zalenski, MD
Jan Penney, RN, MSN
Alan J. Tiefenbrunn, MD
Philip Greenland, MD
for the NRMI Investigators

**Context** Few studies have examined the association between the number of coronary heart disease risk factors and outcomes of acute myocardial infarction in community practice.

**Objective** To determine the association between the number of coronary heart disease risk factors in patients with first myocardial infarction and hospital mortality.

**Design** Observational study from the National Registry of Myocardial Infarction, 1994-2006.

**Patients** We examined the presence and absence of 5 major traditional coronary heart disease risk factors (hypertension, smoking, dyslipidemia, diabetes, and family history of coronary heart disease) and hospital mortality among 542 008 patients with first myocardial infarction and without prior cardiovascular disease.

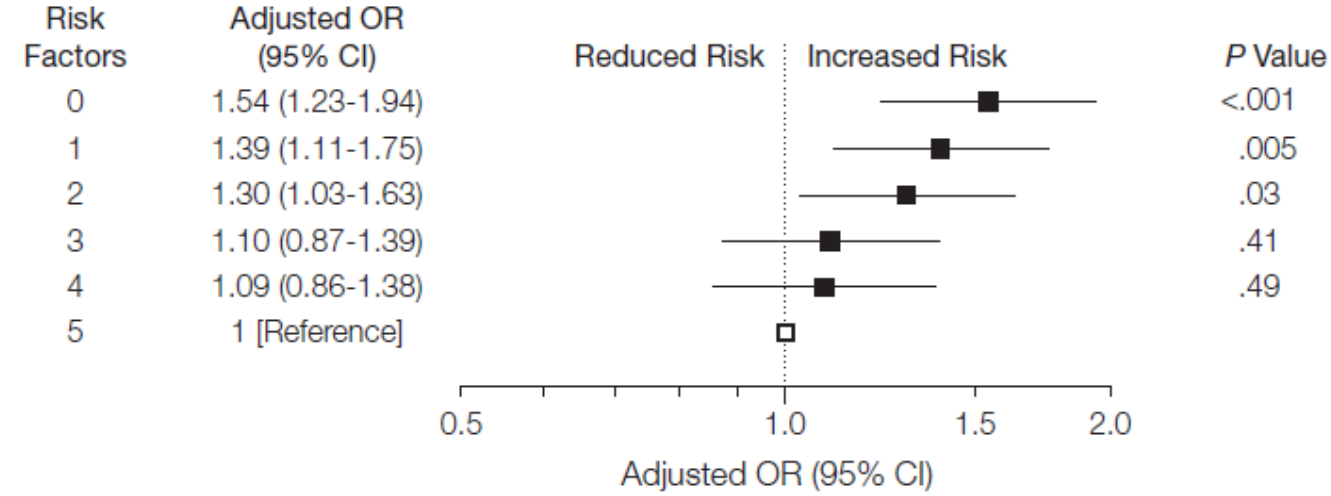
**Results** A majority (85.6%) of patients who presented with initial myocardial infarction had at least 1 of the 5 coronary heart disease risk factors, and 14.4% had none of the 5 risk factors. Age varied inversely with the number of coronary heart disease risk factors, from a mean age of 71.5 years with 0 risk factors to 56.7 years with 5 risk factors ($P$ for trend $<.001$). The total number of in-hospital deaths for all causes was 50 788. Unadjusted in-hospital mortality rates were 14.9%, 10.9%, 7.9%, 5.3%, 4.2%, and 3.6% for patients with 0, 1, 2, 3, 4, and 5 risk factors, respectively. After adjusting for age and other clinical factors, there was an inverse association between the number of coronary heart disease risk factors and hospital mortality adjusted odds ratio (1.54; 95% CI, 1.23-1.94) among individuals with 0 vs 5 risk factors. This association was consistent among several age strata and important patient subgroups.

**Conclusion** Among patients with incident acute myocardial infarction without prior cardiovascular disease, in-hospital mortality was inversely related to the number of coronary heart disease risk factors.

JAMA. 2011;306(19):2120-2127 www.jama.com

# Motivating Example

**Figure 2.** Mortality Risk of Patients With and Without Cardiovascular Risk Factors and First Myocardial Infarction




Conclusion: Having 0 risk factors results in a 54% ***lower*** mortality risk than having 5 risk factors

# Huh?

Should we all be trying to accumulate more CVD risk factors?

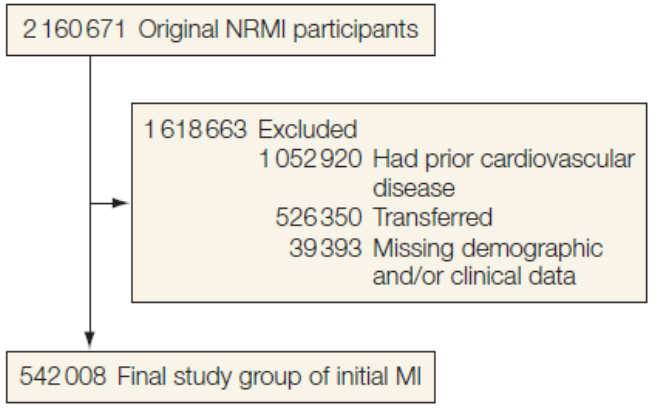
Does this indicate the presence of “novel but as yet uncharacterized and deadly CVD risk factors?”

**Two clues:**

1. Approximately 30% of myocardial infarctions (MIs) lead to death prior to hospitalization

2. The authors report that 75% of individuals who were admitted to the hospital were excluded due to “prior CVD diagnoses”

**Figure 1.** Total NRMI Population, Exclusions, and the Final Study Population

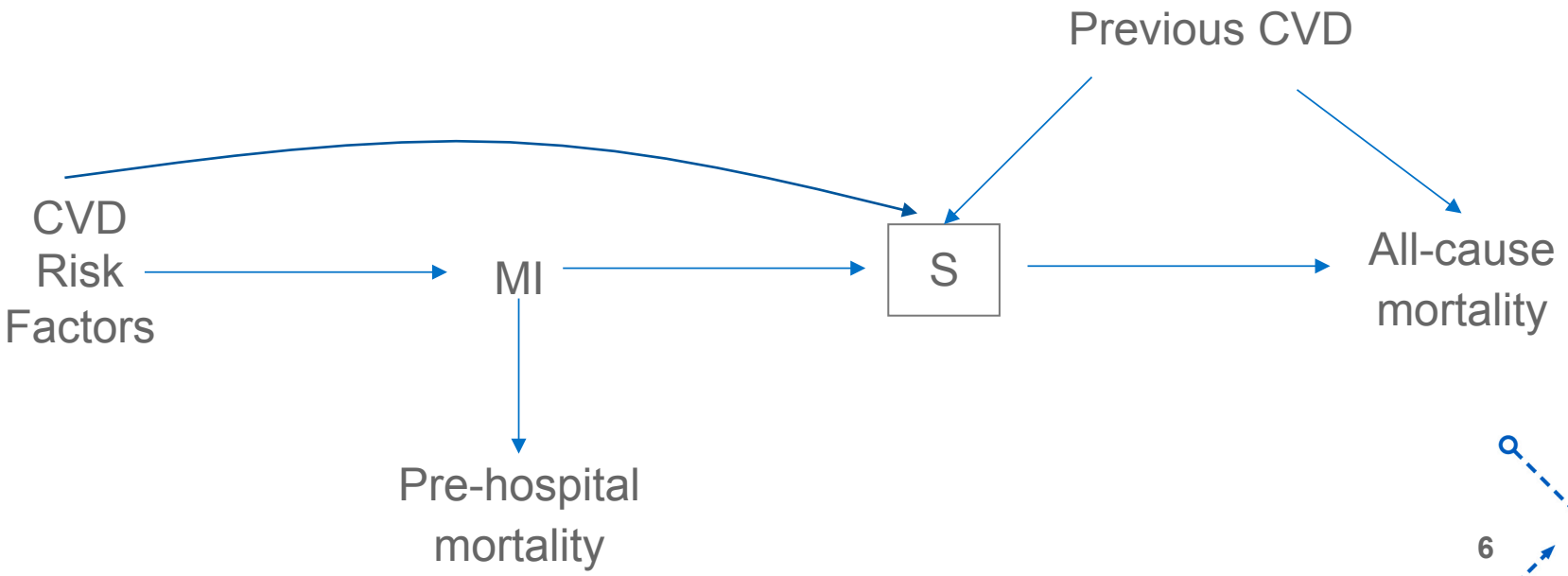



NRMI indicates National Registry of Myocardial Infarction. Prior cardiovascular disease included previous MI, coronary heart disease, angina, heart failure, percutaneous coronary intervention, coronary artery bypass surgery, stroke, cerebrovascular disease, and peripheral vascular disease.

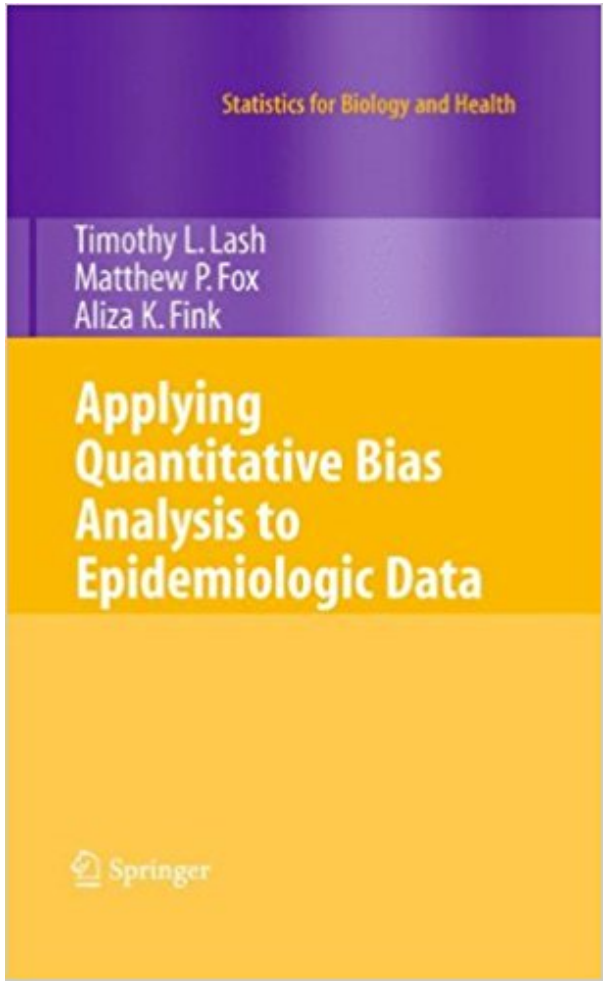
# Selection Bias

- Inclusion in analytic sample was conditional on exposure (number of CVD risk factors) and outcome (mortality) because deaths occurring before hospitalization and patients with existing CVD diagnoses were excluded.

- Individuals with more CVD RF were more likely to be excluded
  - If having more risk factors accelerates mortality (i.e., more risk factors=more likely to die prior to hospitalization) risk factors will appear protective during or after hospitalization



# Disclosure




Tim Lash

Jay Kaufman

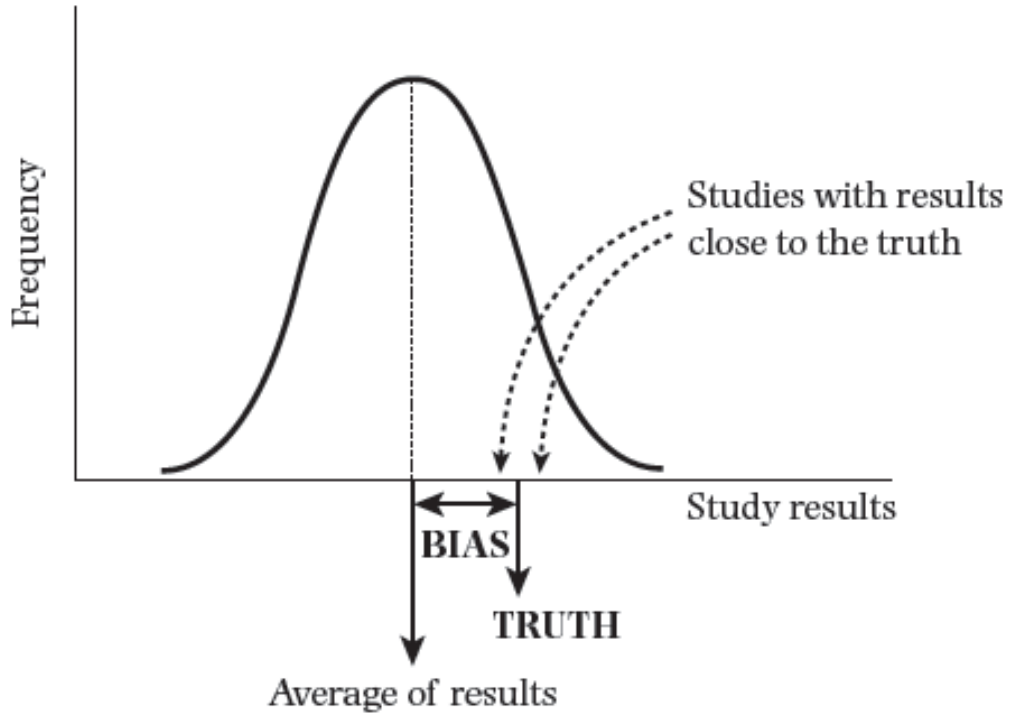
Matt Fox

# Key Terms

## What is bias?

"Systematic deviation of results or inferences from the truth"

**FIGURE 4-1** Hypothetical distribution of results from a biased study design.



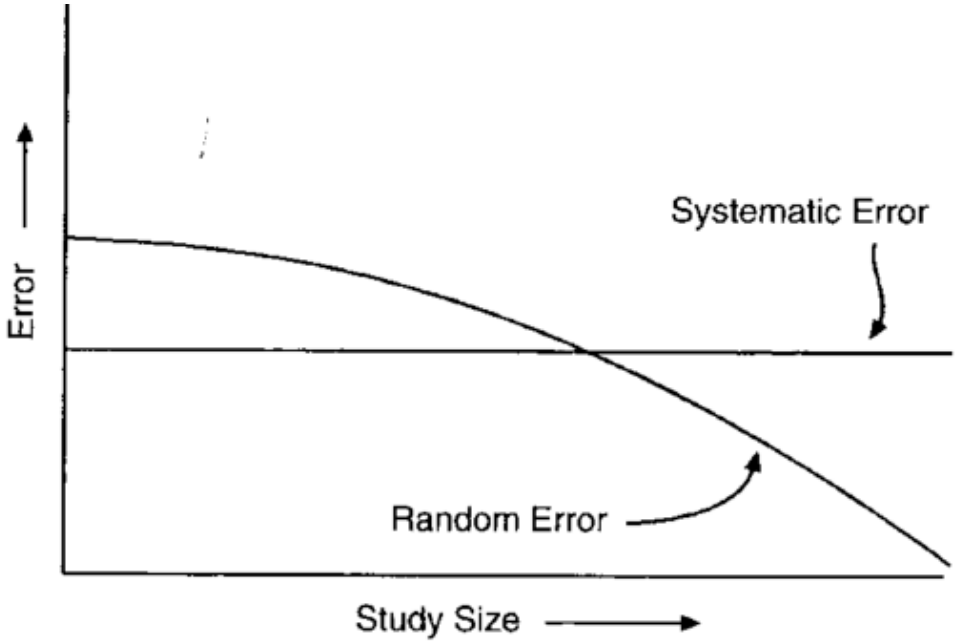

Porta, 2008; Szklo & Nieto,2012

# Key Terms

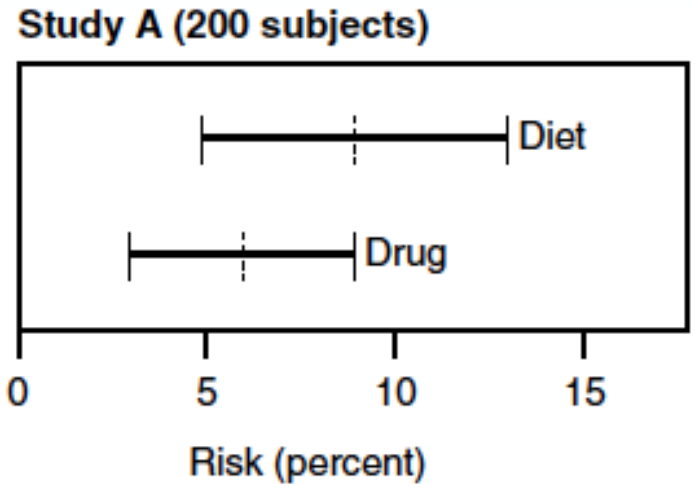
Systematic error vs. random error

Systematic error = bias/lack of validity

Random error = lack of precision



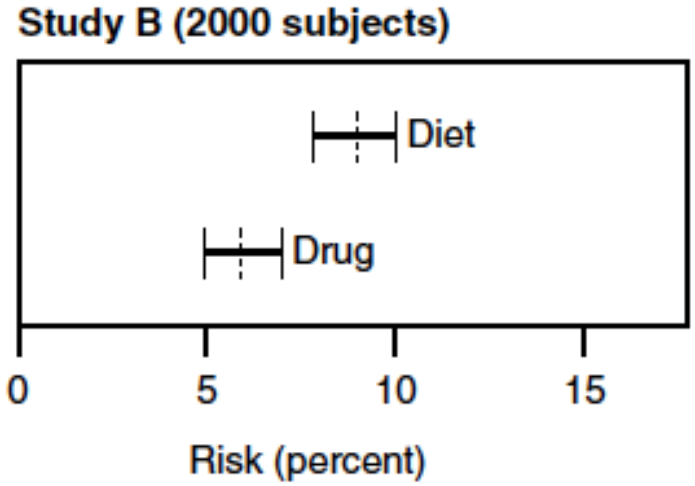
# Random Error




Two hypothetical RCTs

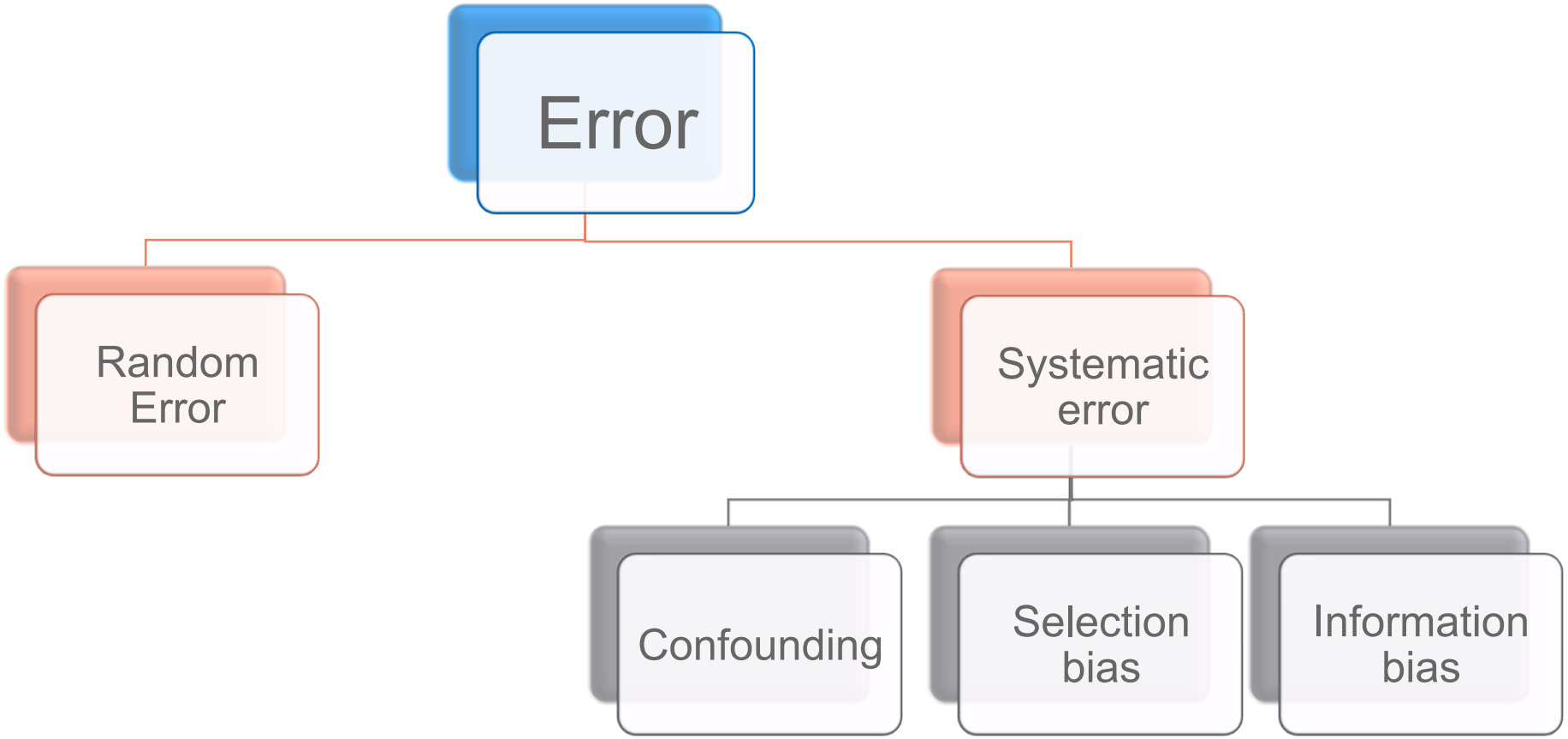
- Cholesterol-lowering drug vs
- Dietary intervention

In both studies, the risk of myocardial infarction (MI) was **9% for the diet** group and **6% for the drug group**




**Study A:** "No difference between treatments"

**Study B:** "Drug is more effective for reducing risk of MI"




**Precision**
Relative lack of random error

**Validity**
Relative absence of systematic error

# Threats to validity

**Selection Bias:**

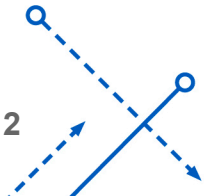
- Baseline
- Losses to follow-up

**Information Bias:**

- Misclassification of variables
- Differential vs. non-differential

**Confounding**

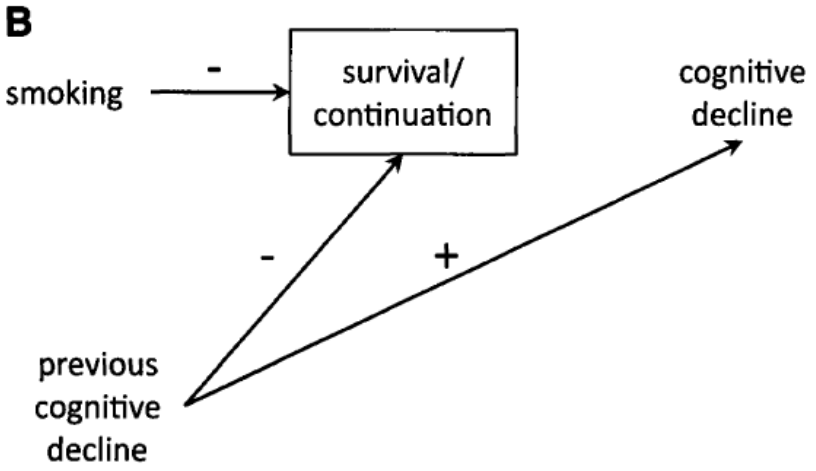
- Unmeasured confounder
- Unknown confounder

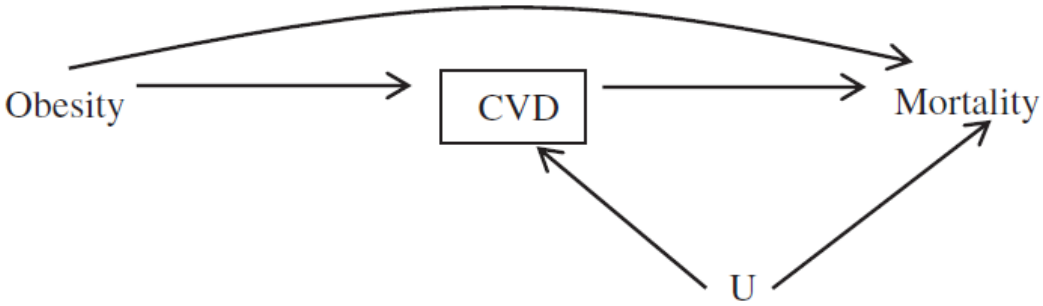


# Selection Bias

- Estimate of association is different in study population than in source population
- Arises when both exposure and outcome affect participation in study

- Collider bias

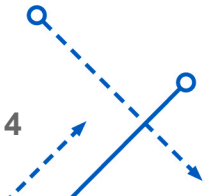


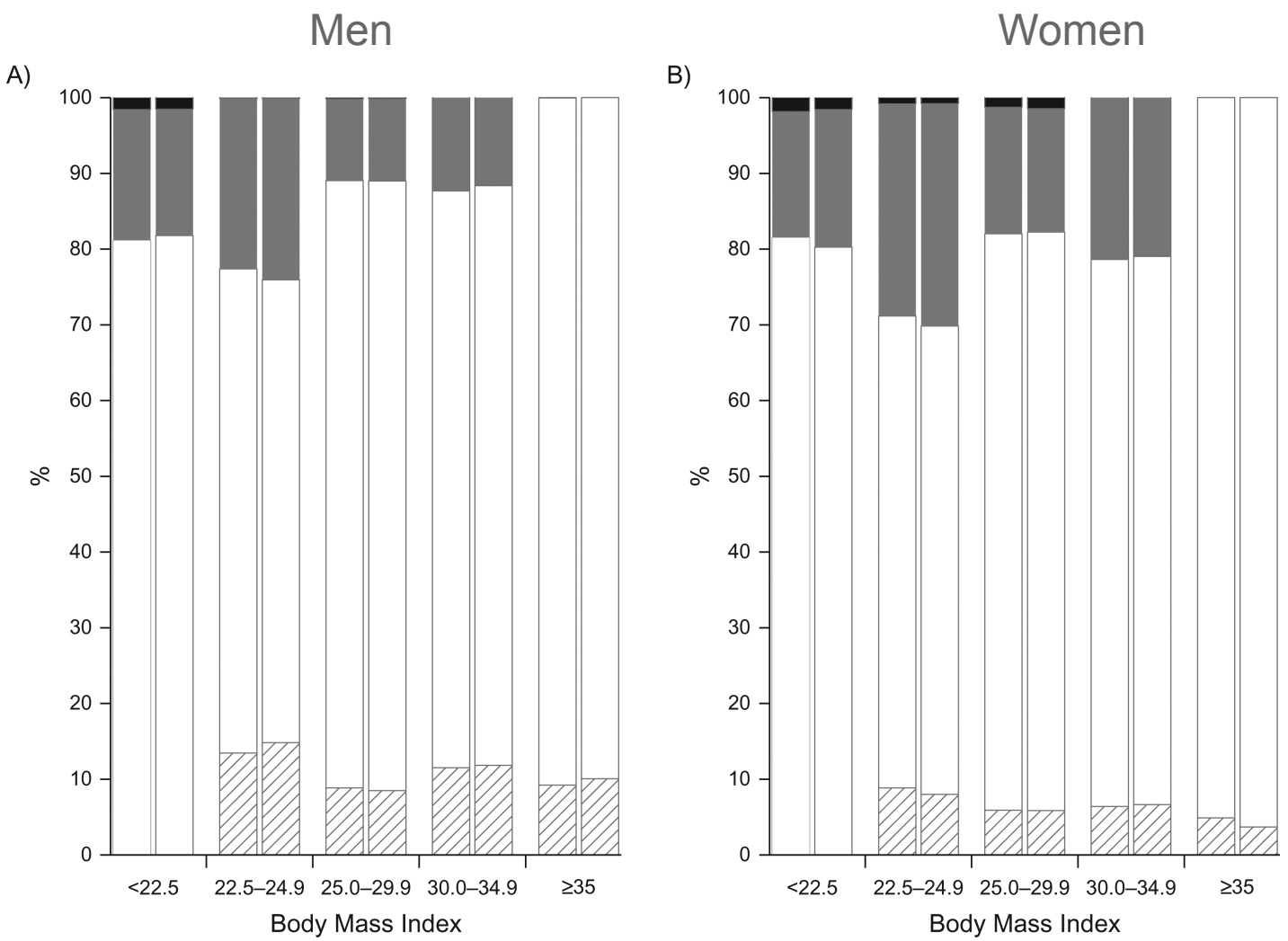




(Banack & Kaufman, 2014; Hernan et al., 2004; Weuve et al., 2012)

# Information Bias

- The estimate of association is distorted by inaccurate measurement or classification of the exposure, outcome, or a covariate.
- Misclassification of categorical variables (exposure, outcome, covariate)
- Measurement error (continuous variables)






White=same
Striped=measured -1
Gray= measured +1
Black= measured +2

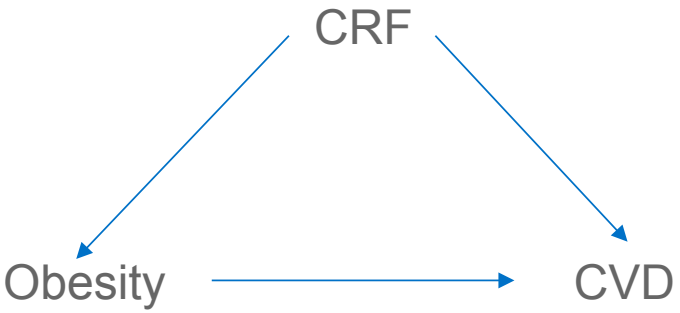
Flegal et al., 2017

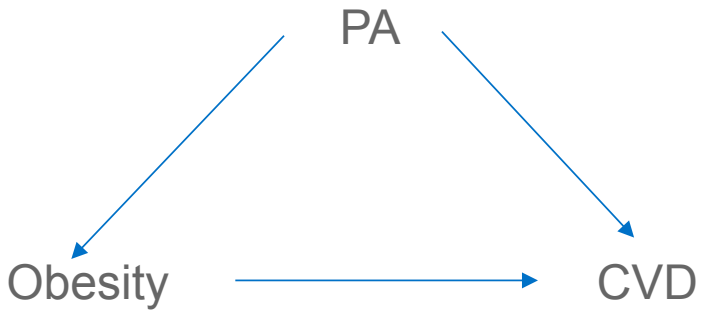
# Confounding

**Confounding**

- Estimate of effect of exposure on outcome is ‘mixed with’ the effect of effect of uncontrolled/unmeasured covariate on outcome

- Unmeasured confounder  (or mismeasured)
- Unknown confounder



# Reliability and validity


| Consistent but wrong | Right answer, on average | Wrong and variable | Consistent and correct |
|---|---|---|---|

# Limitations of meta-analysis

- If there is a systematic bias affecting study results, a meta analysis will not solve this problem
- More precise, but still biased effect estimate

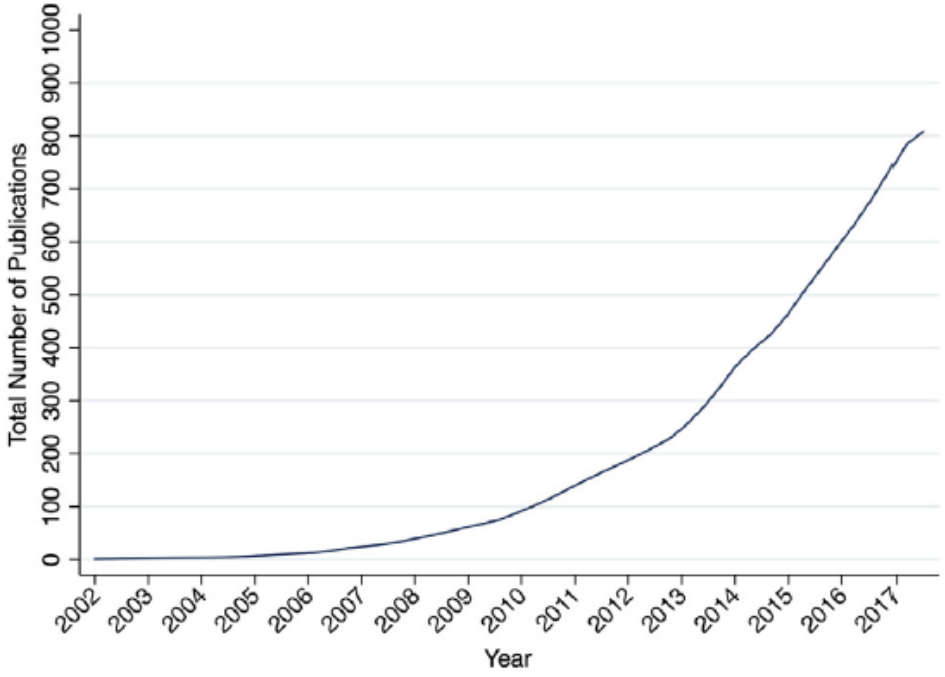



**Figure 1.** Peer-reviewed publications emphasizing the "obesity paradox." Methods included a search of EMBASE (1947-2017, week 29) for full articles that included the phrase "obesity paradox" in the title, abstract, or keywords. Our search strategy identified 812 articles, with the initial articles appearing in the year 2002.

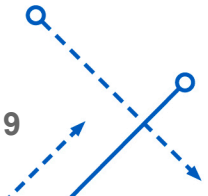
(Stovitz, Banack & Kaufman, 2018)

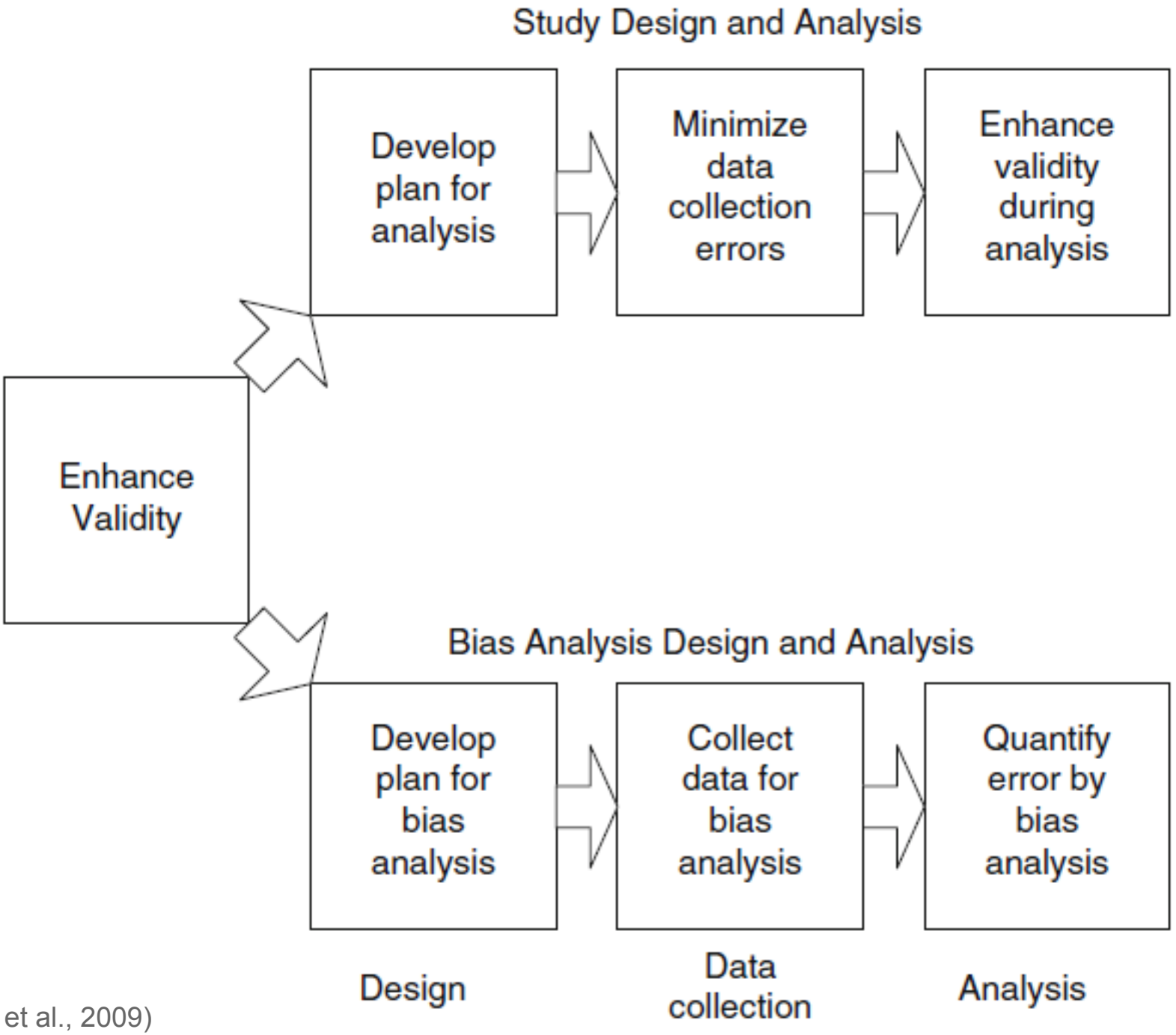
# Why QBA?

As epidemiologists, we all try to design quality studies and reduce the amount of systematic error affecting an effect estimate

**We have a problem, though:**
It's just not always possible

Incumbent upon investigators to quantify how far an estimate is from the "truth" (an estimate that is free of bias)





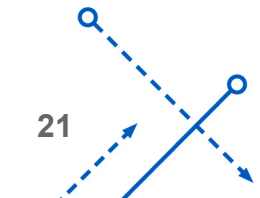

(Lash et al., 2009)

# When QBA?

1. **When the effect estimate is reasonably precise**
    - Narrow 95% CI reflects a small amount of random error
    - With a wider CI, it is already difficult to make inferences because of the amount of random error, regardless of whether you try to account for systematic error

2. **When the effect estimate is affected by a limited number of systematic errors**
    - With multiple systematic errors, total amount of error may be too large to reliably quantify
    - QBA may not be productive; perhaps data should be used for generating ideas for future studies

3. **When you have a way of realistically estimating bias parameters**
    - Required to complete QBA
    - Could come from internal or external validation data
    - Literature review



# Bias parameters

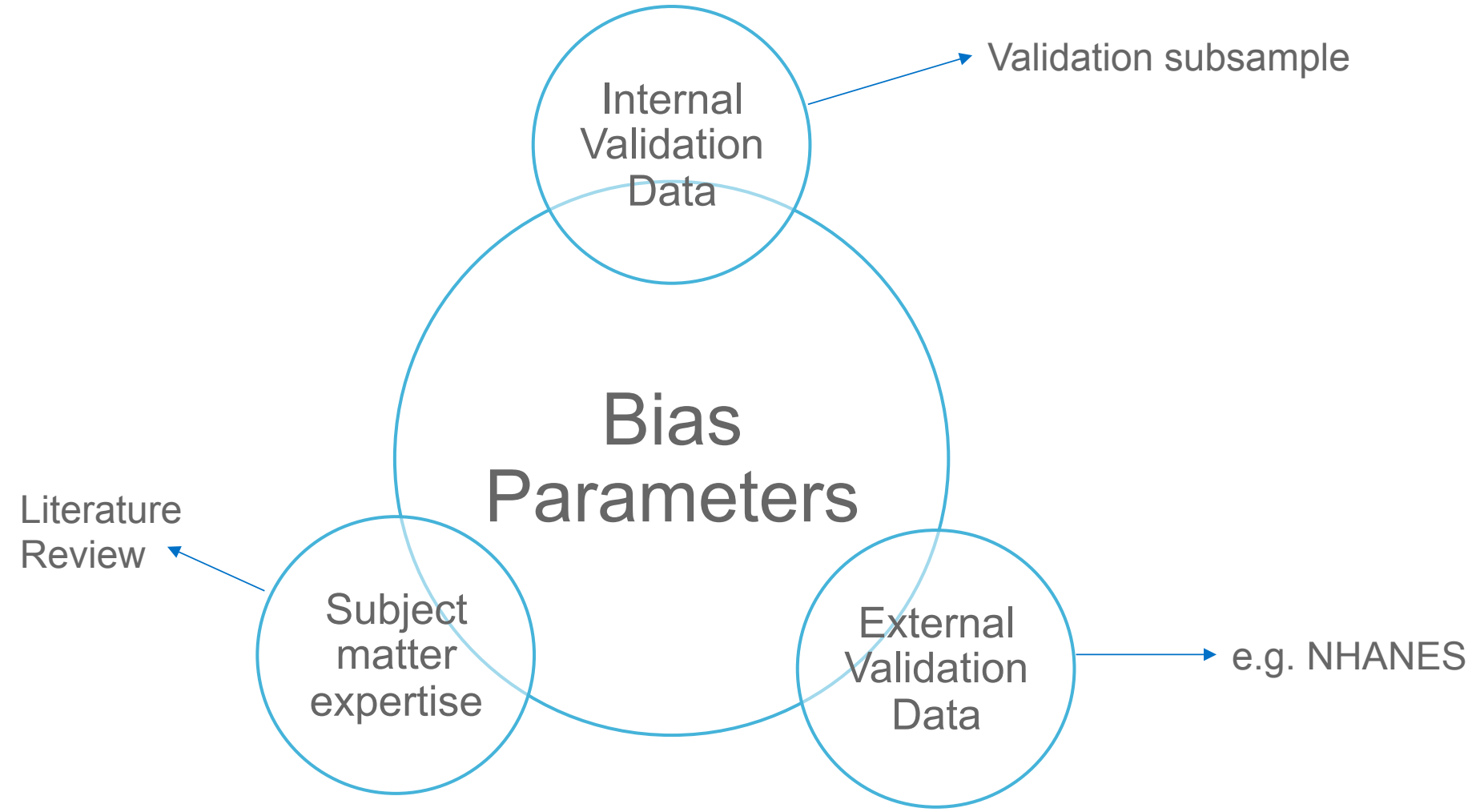
- In QBA, equations are used to adjust an effect estimate to see the impact of potential bias
- The parameters in these equations are called ‘bias parameters’
- Used to determine the direction and magnitude of bias adjustment

Selection bias
-Sampling fractions: proportion of all eligible subjects who participate in your study

Misclassification
-Sensitivity and specificity of exposure classification (or confounder/outcome classification)

Confounding bias
-Hypothesized strength of relationship between unmeasured confounder and exposure and unmeasured confounder and outcome

# Bias parameters



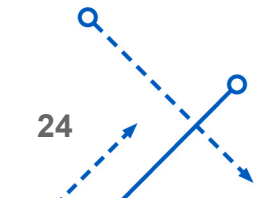
# QBA Techniques

**Deterministic bias analysis:**

- Simple sensitivity analysis
- One fixed value assigned to each bias parameter
- Results in single revised estimate of association

**Probabilistic bias analysis:**

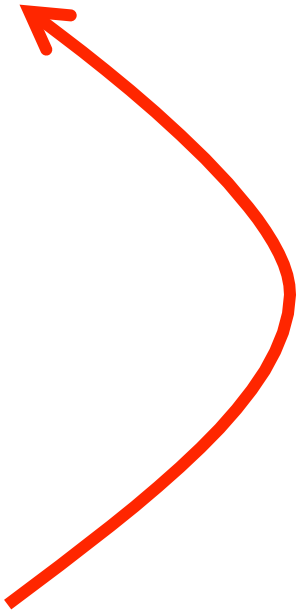
- Probability distribution assigned to bias parameters
- Can account for one bias at a time or multiple biases at a time
- Results in frequency distribution of revised estimates of association
- Computationally intensive



# Multiple Bias Analysis

**When you are dealing with multiple biases affecting a study result, correct for them in the reverse of the order in which they occurred:**

1. • Confounding
2. • Selection Bias
3. • Information Bias

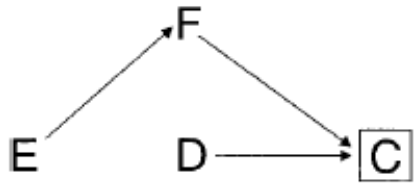
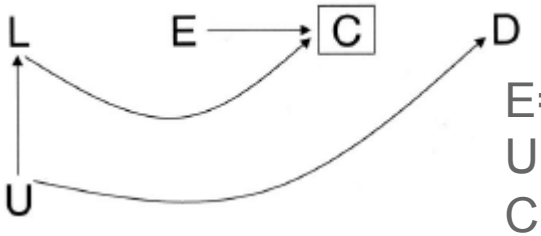
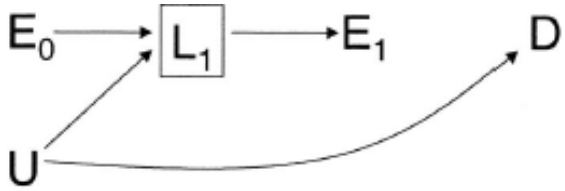


University at Buffalo The State University of New York

# SELECTION BIAS

# Structural Approach to Selection Bias

(Hernan, 2004)

| | | |
|---|---|---|
| Case Control Study<br>*Inappropriate control selection* |  | E=coffee, D=pancreatic cancer, C=selection into study, F=ulcer patients |
| Cohort Study<br>*Differential loss to follow-up* |  | E=HAART, D=AIDS , U=immunosupression, C=censoring L= symptoms (fever, weight loss etc.) |
| Cohort Study<br>*Adjustment for variables affected by previous exposure* |  | E=obesity<br>L=CVD<br>U= genetic factors<br>D=mortality |

# Selection vs. selection bias

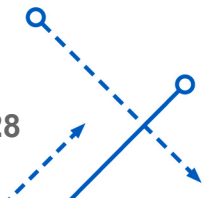
| hbp | final mortality status: Alive | Deceased | Total |
|---|---|---|---|
| normotensive | 6,526 | 246 | 6,772 |
| hypertensive | 2,656 | 531 | 3,187 |
| Total | 9,182 | 777 | 9,959 |

**OR=5.3**

**If you select 50% of cases and 20% of controls- no selection bias:**

| hbp | final mortality status: Alive | Deceased | Total |
|---|---|---|---|
| normotensive | 3,263 | 49 | 3,312 |
| hypertensive | 1,328 | 106 | 1,434 |
| Total | 4,591 | 155 | 4,746 |

**OR=5.3**



# When does selection bias occur?

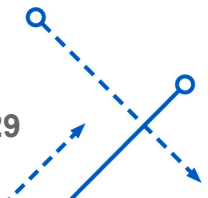
| hbp | final mortality status: Alive | Deceased | Total |
|---|---|---|---|
| normotensive | 6,526 | 246 | 6,772 |
| hypertensive | 2,656 | 531 | 3,187 |
| Total | 9,182 | 777 | 9,959 |

**OR=5.3**

**If you select 50% of unexposed cases, 80% of exposed cases, 20% of unexposed controls, and 10% of exposed controls:**

| hbp | final mortality status: Alive | Deceased | Total |
|---|---|---|---|
| normotensive | 3,263 | 49 | 3,312 |
| hypertensive | 2,125 | 53 | 2,178 |
| Total | 5,388 | 102 | 4,746 |

**OR=1.6**



# Calculate the cross product

**If you select 50% of cases and 20% of controls- no selection bias:**

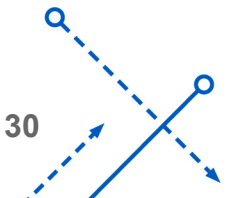
| | Cases | Controls |
|---|---|---|
| Exposed | .5 | .2 |
| Unexposed | .5 | .2 |

(.5*.2) / (.5*.2) = 1

**If you select 50% of unexposed cases, 80% of exposed cases, 20% of unexposed controls, and 10% of exposed controls:**

| | Cases | Controls |
|---|---|---|
| Exposed | .8 | .1 |
| Unexposed | .5 | .2 |

(.8*.2)/(.5*.1) = 0.31



# Sampling Fractions (selection probabilities)

**Target**

| | $E$ | $\bar{E}$ |
|---|---|---|
| $D$ | $A$ | $B$ |
| $\bar{D}$ | $C$ | $D$ |

**Selected Sample**

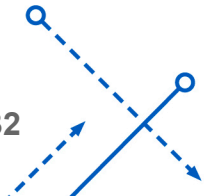
| | $E$ | $\bar{E}$ |
|---|---|---|
| $D$ | $A^{o}$ | $B^{o}$ |
| $\bar{D}$ | $C^{o}$ | $D^{o}$ |

| | |
|---|---|
| $\alpha = A^{\circ} / A$ | $\beta = B^{\circ} / B$ |
| $\gamma = C^{\circ} / C$ | $\delta = D^{\circ} / D$ |

# Bias Analysis for Selection Bias

-In the previous exercise with high blood pressure and mortality, we had the “master” 2x2 table and applied sampling fractions to get a biased table (and biased effect estimate)

-In general, for bias analysis we need to work backwards, we have the biased table (and biased effect estimate), and need to get back to the master 2x2



# Bias analysis for selection bias

1. Need estimate of sampling fractions

| Selection probabilities | Exposure = 1 | Exposure = 0 |
|---|---|---|
| Cases | $S_{\text{case},1}$ | $S_{\text{case},0}$ |
| Noncases | $S_{\text{control},1}$ | $S_{\text{control},0}$ |

2. Use sampling fractions to adjust the biased estimates

$$\text{OR}_{\text{adj}} = \hat{\text{OR}} \times \frac{S_{\text{case},0} S_{\text{control},1}}{S_{\text{case},1} S_{\text{control},0}}$$

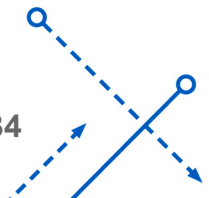
# Example

From a study examining the relationship between obesity and mortality in HF patients:

Proportion

| | Exposed | Unexposed | Total | Exposed |
|---|---|---|---|---|
| Dead | 998 | 977 | 1975 | 0.5053 |
| Alive | 2086 | 1606 | 3692 | 0.5650 |
| Total | 3084 | 2583 | 5667 | 0.5442 |
| | Point estimate | | [95% Conf. | Interval] |
| Odds ratio | .7864429 | | .7037068 | .8789046 |

You suspect there is collider stratification bias due to conditioning on a variable affected by exposure (HF) and sharing common causes with the outcome



# Where to get estimates of sampling fractions?

Need population representative data to estimate selection probabilities

Use data from the 1999-2000 and 2001-2002 US National Health and Nutrition Examination Survey (NHANES)

| | **25.5-29.9** | **18.5-24.9** |
|---|---|---|
| Dead | 258 | 256 |
| Alive | 3270 | 3096 |

Total NHANES cohort

| | **25.5-29.9** | **18.5-24.9** |
|---|---|---|
| Dead | 34 | 29 |
| Alive | 68 | 23 |

NHANES HF=1

| | **25.5-29.9** | **18.5-24.9** |
|---|---|---|
| Dead | 0.132 | 0.113 |
| Alive | 0.021 | 0.007 |

Sampling fractions

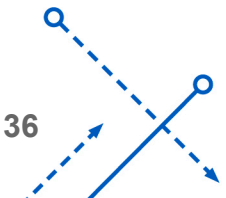
# Calculations

| | 25.5-29.9 | 18.5-24.9 |
|---|---|---|
| Dead | 0.132 | 0.113 |
| Alive | 0.021 | 0.007 |

Sampling fractions

| | 25.5-29.9 | 18.5-24.9 |
|---|---|---|
| Dead | 998 | 977 |
| Alive | 2086 | 1606 |

Biased 2x2 table



# Stata -episens-

- Nicola Orsini and colleagues created a Stata package for bias analysis called **-episens-**
- Add on to Stata (type: findit episens)

- http://www.stata-journal.com/article.html?article=st0138

- Can be used with data or as an immediate command (-episensi-)

```
episensi #a #b #c #d [, options]

Selection bias options
dpscex define the selection probability among cases exposed
dpscun define the selection probability among cases unexposed
dpsnex define the selection probability among noncases exposed
dpsnun define the selection probability among noncases unexposed
```

# -episensi-

```
. episensi 998 977 2086 1606, dpscex(c(.132))
dpscun(c(.113)) dpsnex(c(.0208)) dpsnun(c(.00743))

Pr Case Selection Exposed: Constant(.132)
Pr Case Selection No-Exposed: Constant(.113)
Pr No-Case Selection Exposed: Constant(.0208)
Pr No-Case Selection No-Exposed: Constant(.00743)

Observed Odds Ratio [95% CI]= 0.79 [0.70, 0.88]

Deterministic sensitivity analysis for selection bias
   External adjusted Odds Ratio = 1.88
   Percent bias = -58%
```

# Excel option

Tim Lash et al., have spreadsheets to correct for bias available online as an adjunct to their bias analysis textbook:

https://sites.google.com/site/biasanalysis/

**SELECTION BIAS** — **Chapter 4**

This spreadsheet can be used to conduct a simple sensitivity analysis to correct for selection bias using estimates of the selection proportions. The example follows the example in chapter

Reset Example | Clear Da

**Input Bias Parameters**

| | |
|---|---|
| S(Dead+\|BMI>25+) | 0.13 |
| S(Dead+\|BMI>25-) | 0.11 |
| S(Dead-\|BMI>25+) | 0.02 |
| S(Dead-\|BMI>25-) | 0.01 |
| Error Check: | No errors found |

| Variable Names | |
|---|---|
| Exposure | BMI>25 |
| Outcome | Dead |

| Selection OR |
|---|
| 0.39 |

**Instructions**

Enter the bias parameters in the blue cells to the left and the crude data in the blue cells below. The green cells give the results after adjusting for the selection proportions. Note that white cells are expected values and therefore do not have to be integers.

**Data (Enter Observed Dead-BMI>25 Data in Blue Cells)**

| | Observed Data | | Missing Data | | Corrected for Selection Proportions | |
|---|---|---|---|---|---|---|
| | BMI>25 + | BMI>25 - | BMI>25 + | BMI>25 - | BMI>25 + | BMI>25 - |
| Dead + | 136 a | 107 b | 894.3 $A_1$ | 839.9 $B_1$ | 1030.3 $A_0$ | 946.9 $B_0$ |
| Dead - | 297 c | 165 d | 13845.9 $C_1$ | 23406.4 $D_1$ | 14142.9 $C_0$ | 23571.4 $D_0$ |
| Total | 433 m | 272 n | 14740.2 $M_1$ | 24246.3 $N_1$ | 15173.2 $M_0$ | 24518.3 $N_0$ |

**Observed and Selection Bias Corrected Measures of Dead-BMI>25 Relationship**

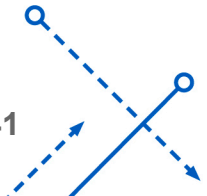
| Observed | Measure (95% CI) | Missing Data Stratum | | Selection Bias Corrected | |
|---|---|---|---|---|---|
| RR (Dead-BMI: | 0.8 (0.65 - 0.98) | RR (Dead-BMI>25) | 1.75 | RR (Dead-BMI>25) | 1.76 |
| OR (Dead-BMI: | 0.71 (0.51 - 0.97) | OR (Dead-BMI>25) | 1.8 | OR (Dead-BMI>25) | 1.81 |

University at Buffalo The State University of New York

# UNMEASURED CONFOUNDING

# Identifying confounding

- Confounding refers to a theoretical concept
  - Risk of disease in the reference population equals the risk the index population would have had, ***if they had not been exposed***
  - Reference group intended to be a substitute for index group in every way ("exchangeable" groups)
  - When they are not exchangeable, there is confounding present
- Methods for accounting for confounding
  - Stratification, MH adjustments, regression modeling



# Crude vs. adjusted estimates

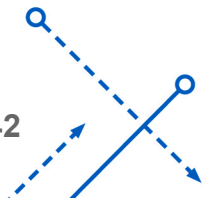
- In practice, confounding typically identified when crude estimate of association does not equal the adjusted estimate (‘change in estimate’)

- Some problems:
  - Only can adjust for variables you have collected information on

  - Can not adjust for:
    - Unknown confounding (variables you can’t identify, theoretical concepts)

    - Unmeasured confounding (variables you did not measure)

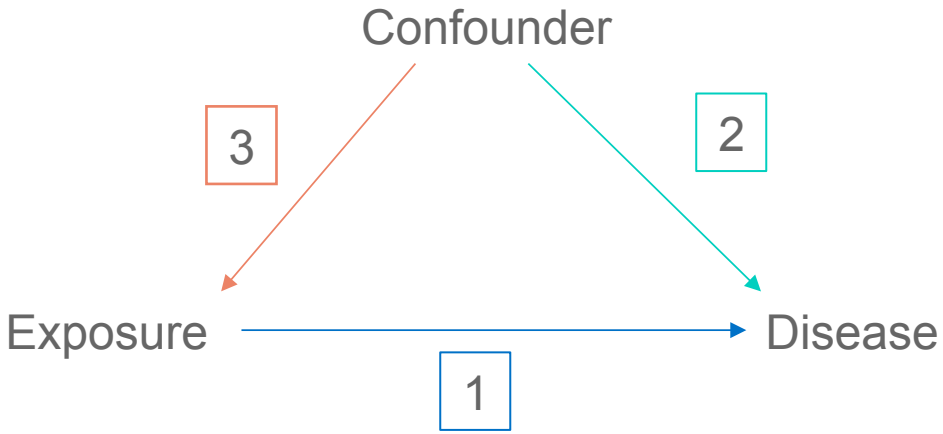
    - Residual confounding (variables measured poorly or by proxy)



# Bias analysis for unmeasured confounding

Requires:

1. Crude estimate of the exposure-outcome association
2. Association between confounder and disease
3. Association between confounder and exposure (or exposure specific prevalence of confounder)




**Assumption:**
Unmeasured confounder is independent of measured confounders

**Result:** Overestimate magnitude of bias

# Question of interest

What estimate of association between EXPOSURE and DISEASE would have been observed in the study had the authors collected data on CONFOUNDER and adjusted for it in the analysis?

$$RR_{adj} = RR_{obs} \frac{RR_{CD}p_0 + (1 - p_0)}{RR_{CD}p_1 + (1 - p_1)}$$

Prevalence of confounder in unexposed (E=0)

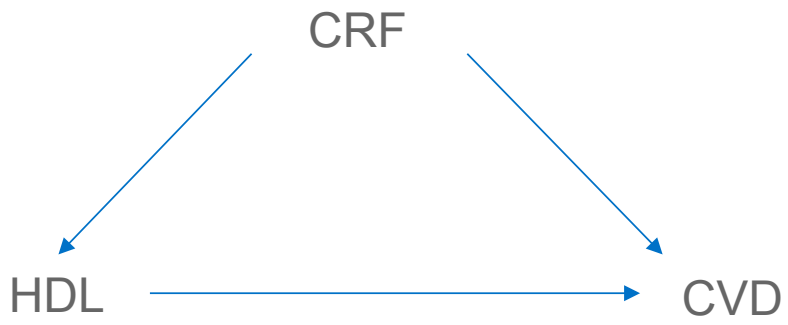
Observed RR without adjustment for confounder

Confounder-disease RR

Prevalence of confounder exposed(E=1)

# Example: HDL, exercise & CVD risk

What estimate of association between HDL and CVD would have been observed in the study had the authors collected data on EXERCISE and adjusted for it in the analysis?



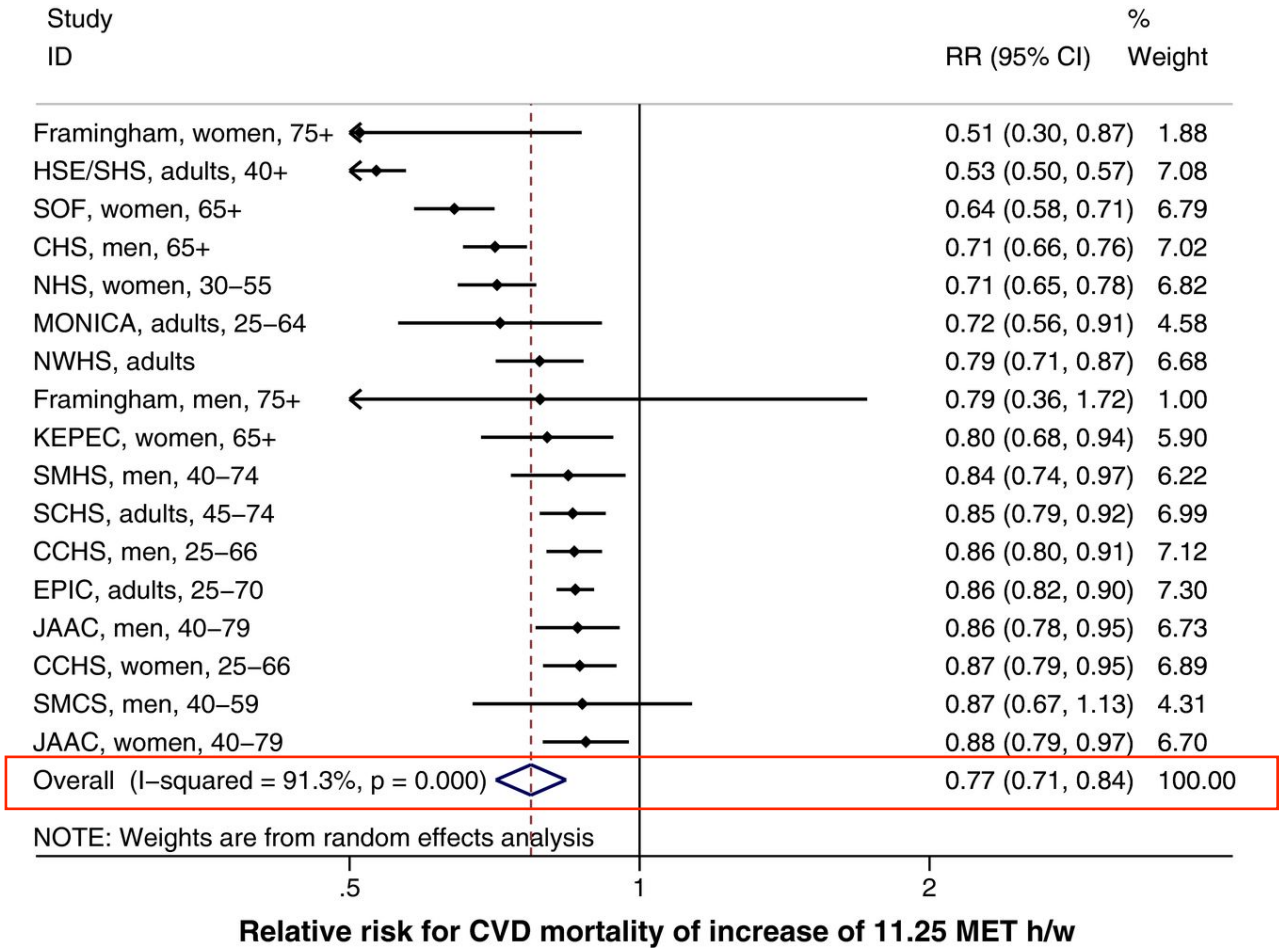
# Crude association HDL → CVD

| | Exposed | Unexposed | Total |
|---|---|---|---|
| Cases | 515 | 4357 | 4872 |
| Noncases | 1894 | 10222 | 12116 |
| Total | 2409 | 14579 | 16988 |
| Risk | .2137817 | .2988545 | .2867907 |

| | Point estimate | [95% Conf. | Interval] |
|---|---|---|---|
| Risk difference | -.0850729 | -.1030516 | -.0670941 |
| Risk ratio | .7153369 | .659999 | .7753145 |
| Prev. frac. ex. | .2846631 | .2246855 | .340001 |
| Prev. frac. pop | .0403669 | | |

chi2(1) = 73.15 Pr>chi2 = 0.0000

# Exercise → ? → CVD




Wahid et al., 2016

# Bias Parameters

| Bias Parameter | Description | Value Assigned |
|---|---|---|
| $RR_{CD}$ | Association between exercise and CVD | 0.77 |
| $p_0$ | Prevalence of exercisers among high HDL | 0.8 |
| $p_1$ | Prevalence of exercisers among low HDL | 0.05 |

# Observed Data

| | Exposed | Unexposed | Total |
|---|---|---|---|
| Cases | 515 | 4357 | 4872 |
| Noncases | 1894 | 10222 | 12116 |
| Total | 2409 | 14579 | 16988 |

| | Total | | $C_1$ | | $C_0$ | |
|---|---|---|---|---|---|---|
| | $E_1$ | $E_0$ | $E_1$ | $E_0$ | $E_1$ | $E_0$ |
| $D_+$ | $a$ | $b$ | $A_1$ | $B_1$ | $A_0$ | $B_0$ |
| $D_-$ | $c$ | $d$ | $C_1$ | $D_1$ | $C_0$ | $D_0$ |
| | $m$ | $n$ | $M_1$ | $N_1$ | $M_0$ | $N_0$ |

# Deterministic bias analysis

| | Total | | $C_1$ | | $C_0$ | |
|---|---|---|---|---|---|---|
| | $E_1$ | $E_0$ | $E_1$ | $E_0$ | $E_1$ | $E_0$ |
| $D_+$ | 515 | 4357 | $A_1$ | $B_1$ | $A_0$ | $B_0$ |
| $D_-$ | 1894 | 10222 | $C_1$ | $D_1$ | $C_0$ | $D_0$ |
| | 2409 | 14579 | 1927.2 | 728.9 | 481.8 | 13850.1 |

1. Solve for $M_1$ and $N_1$ using prevalence of confounder in exposed and unexposed:

$M_1$= 2409 * 0.8 = 1927.2
$N_1$= 14579 * 0.05 = 728.9

2. Solve for $M_0$ and $N_0$ by subtraction:

$M_0$= 2409-1927.2 = 481.8
$N_0$= 14579-728.9 =13850.1

# Deterministic bias analysis

| | Total | | $C_1$ | | $C_0$ | |
|---|---|---|---|---|---|---|
| | $E_1$ | $E_0$ | $E_1$ | $E_0$ | $E_1$ | $E_0$ |
| $D_+$ | $a$ | $b$ | $A_1$ | $B_1$ | $A_0$ | $B_0$ |
| $D_-$ | $c$ | $d$ | $C_1$ | $D_1$ | $C_0$ | $D_0$ |
| | $m$ | $n$ | $M_1$ | $N_1$ | $M_0$ | $N_0$ |

3. Use relationship between exercise and CVD (0.77) to fill in the remaining cells:

$$A_1 = \frac{\mathrm{RR}_{\mathrm{CD}} M_1 a}{\mathrm{RR}_{\mathrm{CD}} M_1 + m - M_1} = 0.77 * 1927.2 * 515 / 0.77 * 1927.2 + 2409 - 1927.2 = 388.8$$

$$B_1 = \frac{\mathrm{RR}_{\mathrm{CD}} N_1 b}{\mathrm{RR}_{\mathrm{CD}} N_1 + n - N_1} = 0.77 * 728.9 * 4357 / 0.77 * 728.9 + 14579 - 728.9 = 169.7$$

# Deterministic bias analysis

| | Total | | $C_1$ | | $C_0$ | |
|---|---|---|---|---|---|---|
| | $E_1$ | $E_0$ | $E_1$ | $E_0$ | $E_1$ | $E_0$ |
| $D_+$ | $a$ | $b$ | $A_1$ | $B_1$ | $A_0$ | $B_0$ |
| $D_-$ | $c$ | $d$ | $C_1$ | $D_1$ | $C_0$ | $D_0$ |
| | $m$ | $n$ | $M_1$ | $N_1$ | $M_0$ | $N_0$ |

4. Fill in remaining cells of the table and calculate effect estimate:

| | E1 | E0 | E1 | E0 | E1 | E0 |
|---|---|---|---|---|---|---|
| D+ | 515 | 4357 | 388.8 | 169.7 | 126.2 | 4187.3 |
| D- | 1894 | 10222 | 1538.4 | 559.3 | 355.6 | 9662.7 |
| | 2409 | 14579 | 1927.2 | 729.9 | 481.8 | 13850.1 |

Corrected RR= 0.83

This spreadsheet can be used to conduct a simple sensitivity analysis to correct for unknown or unmeasured confounding. The example follows chapter 5.

Reset Example | Clear Data

## Instructions

Enter bias parameters in blue cells to the right and the crude data in the blue cells below. Cells in green give the results after adjusting for the unmeasured confounder.

## Input Bias Parameters

| Variable Names | | Bias Parameters | |
|---|---|---|---|
| Outcome | CVD | p(Exercise+\|HDL+) | 0.80 |
| Exposure | HDL | p(Exercise+\|HDL-) | 0.05 |
| Confounder | Exercise | RR(Exercise-CVD) | 0.77 |
| Error Check: | No errors found | | |

## Data (Enter Crude HDL-CVD Data in Blue Cells)

| | Total | | Exercise + | | Exercise - | |
|---|---|---|---|---|---|---|
| | HDL + | HDL - | HDL + | HDL - | HDL + | HDL - |
| CVD + | 515 a | 4357 b | 388.8 $A_1$ | 169.7 $B_1$ | 126.2 $A_0$ | 4187.3 $B_0$ |
| CVD - | 1894 c | 10222 d | 1538.4 $C_1$ | 559.3 $D_1$ | 355.6 $C_0$ | 9662.7 $D_0$ |
| Total | 2409 m | 14579 n | 1927.2 $M_1$ | 729.0 $N_1$ | 481.8 $M_0$ | 13850.1 $N_0$ |

## Crude and Unmeasured Confounder Specific Measures of HDL-CVD Relationship

| Crude | Measure (95% CI) | Exercise + | | Exercise - | |
|---|---|---|---|---|---|
| RR (HDL-CVD) | 0.72 (0.66 - 0.78) | RR (HDL-CVD) | 0.87 | RR (HDL-CVD) | 0.87 |

## HDL-CVD Relationship Adjusted for Exercise

| Standardized Morbidity Ratio | | | Mantel-Haenszel | | |
|---|---|---|---|---|---|
| $SMR_{RR}$ | 0.87 | RRc 0.83 | $MH_{RR}$ | 0.87 | RRc 0.83 |

# -episens- or -episensi-

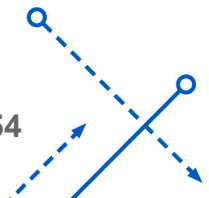
To use -episens- for analysis of unmeasured confounding, need estimates of the bias parameters as previously described:

```
dpexp: define the prevalence of the confounder among E+

dpunexp: define the prevalence of the confounder among E-

drrcd: define the confounder-disease relative risk

dorce: define the confounder-exposure odds ratio
```



# -episens- or -episensi-

```
episensi 515 4357 1894 10222, st(cs) dpexp(c.(0.8)) dpunexp(c.(0.05)) drrcd(c(0.77))

Pr(c=1|e=1): Constant(.8)
Pr(c=1|e=0): Constant(.05)
RR_cd      : Constant(.77)

Observed Risk Ratio [95% Conf. Interval]= 0.72 [0.66, 0.78]

Deterministic sensitivity analysis for unmeasured confounding
   External adjusted Risk Ratio = 0.82
   Percent bias = -13%
```

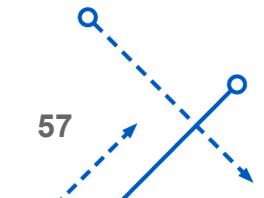
University at Buffalo The State University of New York

# MISCLASSIFICATION

# Two types of misclassification

1. Incorrect measurement and recording of values:
    - Instruments not calibrated correctly
    - Participants not answering questions accurately
    - Incorrect data entry
2. Discordance between definition of variable used in the study and the ‘truth’

**Result:** When dividing participants into exposed/unexposed/diseased/non-diseased some people will be classified into the wrong group (=misclassification)



# Terminology

| | Truly exposed | Truly unexposed |
|---|---|---|
| Classified as exposed | $A$ | $B$ |
| Classified as unexposed | $C$ | $D$ |

$$\text{Sensitivity (SE)} = \frac{A}{A+C}$$

$$\text{Positive predictive value (PPV)} = \frac{A}{A+B}$$

$$\text{Specificity (SP)} = \frac{D}{B+D}$$

$$\text{Negative predictive value (NPV)} = \frac{D}{C+D}$$

Sensitivity= Probability someone is truly exposed is classified as exposed
Specificity= Probability someone is truly unexposed is classified as unexposed

PPV= Probability of truly being exposed if classified as exposed
NPV= Probability of truly being unexposed if classified as unexposed

Predictive values are highly dependent on the true prevalence of exposure in the population in which the predictive value is measured

# Other key terms

- **Non-differential misclassification**: Assignment of exposure is independent of disease status (e.g., cohort study where disease hasn't occurred, blinded ascertainment) OR  assignment of disease is independent of exposure status

- **Differential misclassification:** Sensitivity and/or specificity for disease classification depends on exposure status OR sensitivity and/or specificity for exposure classification depends on disease status

- **Dependent Misclassification:** When individuals who are misclassified on one variable (e.g., exposure status) are more likely than individuals correctly classified on that variable to also be misclassified on a second variable (e.g., disease status)

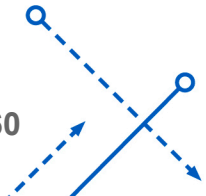
- **Independent Misclassification:** When individuals who are misclassified on one variable are no more likely to be misclassified on a second variable

# Where to obtain bias parameters?

- Internal validation study

| | Strategy | Parameters |
|---|---|---|
| 1 | Select individuals based on misclassified exposure measurement | PPV, NPV |
| 2 | Select individuals based on true exposure status | SE, SP |
| 3 | Select a random sample of individuals | PPV, NPV, SE, SP |

- External sources (i.e., literature review, expert collaboration, educated guesses)



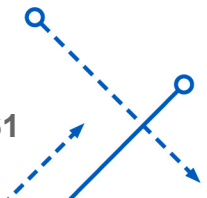
# Example: Self-report vs measured BMI

Measured BMI

| bmi_mcat | Freq. | Percent | Cum. |
|---|---|---|---|
| <18.5 | 646 | 2.01 | 2.01 |
| 18.5-24.9 | 10,140 | 31.55 | 33.56 |
| 25-29.9 | 11,136 | 34.65 | 68.22 |
| 30-34.9 | 6,345 | 19.74 | 87.96 |
| 35-39.9 | 2,758 | 8.58 | 96.55 |
| >40 | 1,110 | 3.45 | 100.00 |
| Total | 32,135 | 100.00 | |

Self-report BMI

| bmi_srcat | Freq. | Percent | Cum. |
|---|---|---|---|
| <18.5 | 579 | 1.81 | 1.81 |
| 18.5-24.9 | 11,345 | 35.50 | 37.32 |
| 25-29.9 | 11,129 | 34.83 | 72.14 |
| 30-34.9 | 5,838 | 18.27 | 90.41 |
| 35-39.9 | 2,204 | 6.90 | 97.31 |
| >40 | 859 | 2.69 | 100.00 |
| Total | 31,954 | 100.00 | |



# Comparing self-report and measured BMI

| bmi_mcat | bmi_srcat | | | | | |
|---|---|---|---|---|---|---|
| | <18.5 | 18.5-24.9 | 25-29.9 | 30-34.9 | 35-39.9 | >40 |
| <18.5 | 346 | 273 | 0 | 0 | 0 | 0 |
| 18.5-24.9 | 194 | 8,758 | 862 | 15 | 0 | 3 |
| 25-29.9 | 8 | 1,923 | 8,206 | 579 | 20 | 2 |
| 30-34.9 | 0 | 92 | 1,730 | 4,047 | 222 | 12 |
| 35-39.9 | 0 | 6 | 102 | 988 | 1,449 | 93 |
| >40 | 0 | 2 | 7 | 68 | 413 | 524 |

There doesn't even appear to be a consistent direction to the misclassification, people are just bad at self-reporting their BMI accurately

**Table 2.** Mortality Hazard Ratios[a] for Measured or Self-Reported Categories of Body Mass Index According to Sex in the Full and Restricted[b] Samples From National Health and Nutrition Examination Survey II (1976–1980), III (1988–1994), and Continuous (1999–2010), United States

| Sample and BMI Ascertainment Method | BMI Category | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Low (<22.5) | | Overweight (25–29.9) | | Class I Obesity (30–34.9) | | Class II–III Obesity (≥35) | |
| | HR | 95% CI | HR | 95% CI | HR | 95% CI | HR | 95% CI |
| *Men* | | | | | | | | |
| Full sample | | | | | | | | |
| Measured[c] | 1.28 | 1.13, 1.45 | 0.94 | 0.86, 1.02 | 1.08 | 0.95, 1.22 | 1.70 | 1.42, 2.03 |
| Self-reported[d] | 1.30 | 1.16, 1.44 | 0.97 | 0.89, 1.05 | 1.07 | 0.94, 1.22 | 1.83 | 1.52, 2.20 |
| Restricted sample | | | | | | | | |
| Measured[c] | 0.87 | 0.56, 1.36 | 0.85 | 0.62, 1.17 | 1.13 | 0.78, 1.64 | 1.86 | 1.12, 3.07 |
| Self-reported[d] | 1.26 | 0.82, 1.94 | 1.24 | 0.90, 1.70 | 1.26 | 0.84, 1.90 | 2.88 | 1.76, 4.72 |

Abbreviations: BMI, body mass index; CI, confidence interval; HR, hazard ratio.

[a] Hazard ratios from Cox proportional hazards models, with age as the timeline and adjusted for survey, smoking status, racial/ethnic group, and alcohol consumption. The reference category was BMI of 22.5–24.9.

[b] The restricted sample was limited to participants who reported never smoking, who reported no history of heart disease or cancer, and who were examined when younger than age 70 years.

[c] Measured BMI was calculated from measured height and measured weight as weight (kg)/height $(m)^2$.

[d] Self-reported BMI was calculated from self-reported weight and self-reported height as weight (kg)/height $(m)^2$.

# Using observed data to calculate the 'truth'

| | Observed | | Corrected data | |
|---|---|---|---|---|
| | $E_1$ | $E_0$ | $E_1$ | $E_0$ |
| $D_+$ | $a$ | $b$ | $[a - D_{+\,\text{Total}}(1 - \text{SP}_{D+})]/[\text{SE}_{D+} - (1 - \text{SP}_{D+})]$ | $D_{+\,\text{Total}} - A$ |
| $D_-$ | $c$ | $d$ | $[c - D_{-\,\text{Total}}(1 - \text{SP}_{D-})]/[\text{SE}_{D-} - (1 - \text{SP}_{D-})]$ | $D_{-\,\text{Total}} - C$ |
| Total | $a + c$ | $b + d$ | $A + C$ | $B + D$ |

$A + C$ is the corrected total number of exposed individuals and $B + D$ is the corrected total number of unexposed individuals.

# Example: Race and obesity

You're conducting a study to explore the relationship between black race (black=exposed, white=unexposed) and obesity (obese=cases, non-obese=non-cases)

All that is available in the dataset is self-report BMI.

| | Exposed | Unexposed | Total |
|---|---|---|---|
| Cases | 2670 | 7957 | 10627 |
| Noncases | 4357 | 18710 | 23067 |
| Total | 7027 | 26667 | 33694 |
| Risk | .379963 | .2983838 | .3153974 |

| | Point estimate | [95% Conf. | Interval] |
|---|---|---|---|
| Risk ratio | 1.273404 | 1.229504 | 1.318871 |

# Validation Study

After taking a QBA workshop, you decide to conduct a validation study to examine potential non-differential misclassification of the outcome in your study (self-report BMI) :

| ob_sr | ob_m 1 | 0 |
|---|---|---|
| 1 | 9,279 | 1,348 |
| 0 | 2,486 | 20,581 |

Calculate sensitivity, specificity, PPV, and NPV from this table

# -Diagt- command

Report summary statistics for diagnostic tests compared to true disease status

---

diagt diagvar testvar [weight] [if exp] [in range] [, prev(#) level(#)

diagti #a #b #c #d [, prev(#) level(#)

**diagvar** is the variable which contains the real status of the patient, and **testvar** is the variable which identifies the result of the diagnostic test.

# -Diagt-

```
. diagt obm obsr

           |         obsr
       obm |      Pos.       Neg. |     Total
-----------+----------------------+----------
  Abnormal |     9,279      2,486 |    11,765
    Normal |     1,348     20,581 |    21,929
-----------+----------------------+----------
     Total |    10,627     23,067 |    33,694

True abnormal diagnosis defined as obm = 1
```

| | | | [95% Confidence Interval] | |
|---|---|---|---|---|
| Prevalence | Pr(A) | 35% | 34% | 35.4% |
| Sensitivity | Pr(+\|A) | 78.9% | 78.1% | 79.6% |
| Specificity | Pr(-\|N) | 93.9% | 93.5% | 94.2% |
| ROC area | (Sens. + Spec.)/2 | .864 | .86 | .868 |
| Likelihood ratio (+) | Pr(+\|A)/Pr(+\|N) | 12.8 | 12.2 | 13.5 |
| Likelihood ratio (-) | Pr(-\|A)/Pr(-\|N) | .225 | .217 | .233 |
| Odds ratio | LR(+)/LR(-) | 57 | 53.1 | 61.2 |
| Positive predictive value | Pr(A\|+) | 87.3% | 86.7% | 87.9% |
| Negative predictive value | Pr(N\|-) | 89.2% | 88.8% | 89.6% |

# Excel spreadsheet

**OUTCOME MISCLASSIFICATION** — **Chapter 6**

This spreadsheet can be used to conduct a simple sensitivity analysis to correct for outcome misclassification. The example follows the example in chapter 6.

Reset Example | Clear Data

**Input Bias Parameters**

| | |
|---|---|
| Se (black+) | 0.79 |
| Se (black-) | 0.79 |
| Sp (black+) | 0.94 |
| Sp (black-) | 0.94 |

Error Check: No errors

**Variable Names**

| | |
|---|---|
| Outcome | obese |
| Exposure | black |

**Instructions**

Enter the bias parameters in the blue cells to the left and the observed data in the blue cells below. Cells in green give the results after correcting for outcome misclassification. Note: green cells are expected values and therefore do not have to be integers.

**Data (Enter black-obese Data in Blue Cells)**

Observed Data

| | black + | black - | Total |
|---|---|---|---|
| obese + | 2670 (a) | 7957 (b) | 10627 |
| obese - | 4357 (c) | 18710 (d) | 23067 |
| Total | 7027 (m) | 26667 (n) | |

| Observed | Measure (95% CI) |
|---|---|
| RR (black-obese) | 1.27 (1.23 - 1.32) |
| OR (black-obese) | 1.44 (1.36 - 1.52) |

**Corrected Data**

Corrected Data

| | black + | black - | Total |
|---|---|---|---|
| obese + | 3078.8 (A) | 8695.5 (B) | 11774.3 |
| obese - | 3948.2 (C) | 17971.5 (D) | 21919.7 |
| Total | 7027 (M) | 26667 (N) | |

| Corrected | Measure |
|---|---|
| RR (black-obese) | 1.34 |
| OR (black-obese) | 1.61 |

**Notes**

**Equations**

$A = [a - E+(1 - SP_{E+})] / [SE_{E+} - (1 - SP_{E+})]$

$B = [b - E-(1 - SP_{E-})] / [SE_{E-} - (1 - SP_{E-})]$

$C = E+ - A$

$D = E- - B$

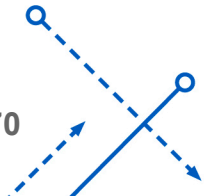
# Using the spreadsheet….

What would the corrected RR be if the sensitivity was 50% and specificity was 80%?

Upon further examination, it appears as though the misclassification of obesity status is differential by exposure status. Using the following parameters, what is the corrected RR?

- Sensitivity among obese black people= 64%
- Sensitivity among non-obese non-black people= 83%

- Specificity among obese black people= 95%
- Specificity among non-obese non-black people= 90%



# -episensi- for misclassification

```
episensi  2670 7957 4357 18710, st(cc) dseca(c(.64)) dspca(c(.
83)) dsenc(c(.95)) dspnc(c(.90))

Se|Cases    : Constant(.64)
Sp|Cases    : Constant(.83)
Se|No-Cases: Constant(.95)
Sp|No-Cases: Constant(.9)

Observed Odds Ratio [95% Conf. Interval]= 1.44 [1.36, 1.52]

Deterministic sensitivity analysis for misclassification of the
exposure
   External adjusted Odds Ratio = 1.79
   Percent bias = -19%
```

# Using PPV and NPV

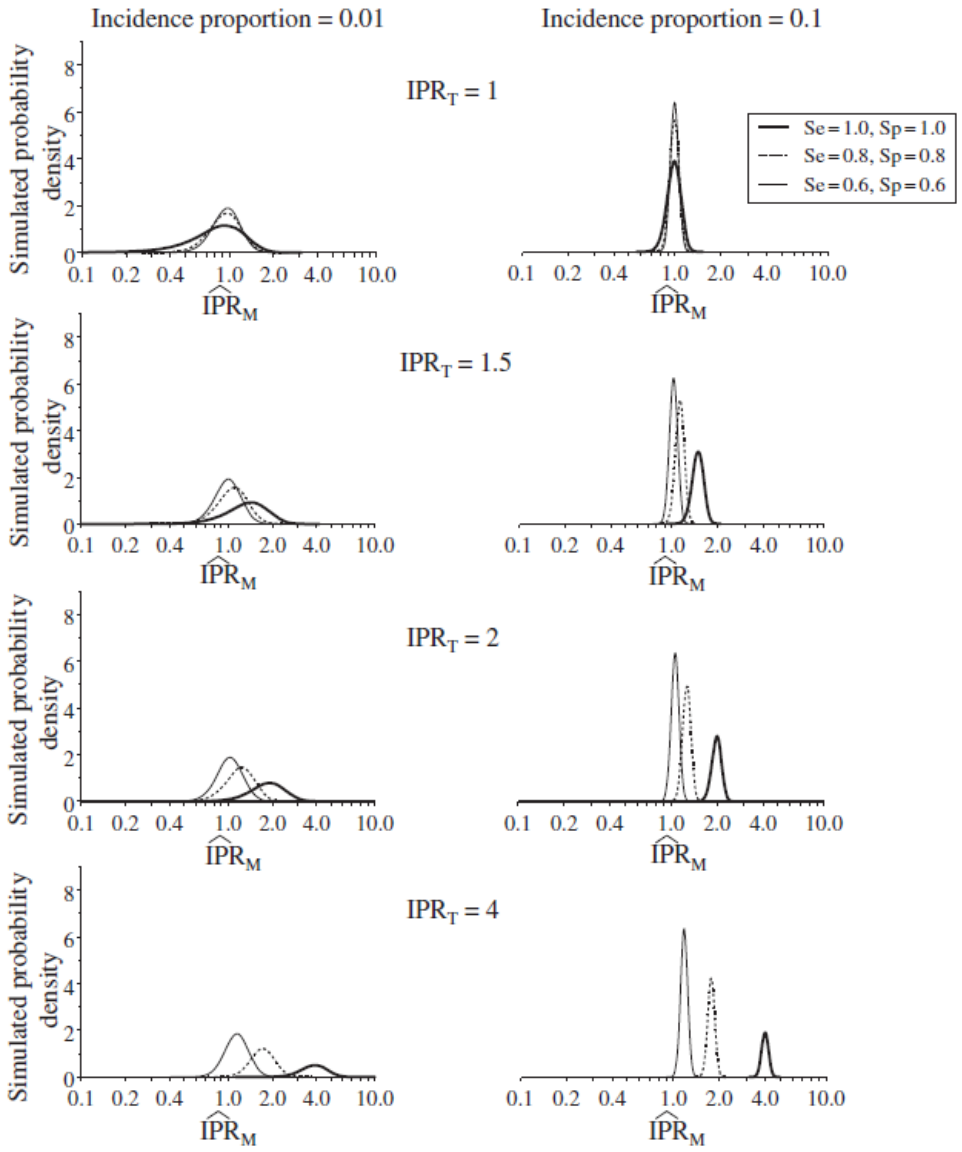
When PPV and NPV data are available, can use those values in place of sensitivity and specificity:

| | Observed | | Corrected data | |
|---|---|---|---|---|
| | $E_1$ | $E_0$ | $E_1$ | $E_0$ |
| $D_+$ | $a$ | $b$ | $a(\text{PPV}_{D+}) + b(1 - \text{NPV}_{D+})$ | $D_{+\,\text{Total}} - A$ |
| $D_-$ | $c$ | $d$ | $c(\text{PPV}_{D-}) + d(1 - \text{NPV}_{D-})$ | $D_{-\,\text{Total}} - C$ |
| Total | $a + c$ | $b + d$ | $A + C$ | $B + D$ |

# Biggest misconception

Non-differential misclassification does not always lead to a bias toward the null!

The simulations by Jurek et al. demonstrate that on average, results may be biased toward the null, but for some individual iterations, the result was biased away from the null




Jurek et al., 2005

# It's not always true

- When an exposure has more than 2 categories, misclassification of exposure can bias estimates of effect in middle categories away from the null

**Table 6.14** Association between body mass index and the risk of diabetes, BRFSS 2001 (Mokdad et al., 2003)

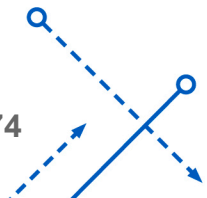
| | <25 | 25–30 | 30+ |
|---|---|---|---|
| Diabetics | 3,053 | 5,127 | 6,491 |
| Nondiabetics | 81,416 | 65,104 | 33,814 |
| OR | Reference | 2.1 | 5.1 |

Validation study demonstrated BRFSS underestimated overweight and obesity

Sp=100%
$Se_{over}$= 91%
$Sp_{over}$= 68.2%

**Table 6.15** Corrected data for the association between body mass index and prevalence of diabetes (adapted from Mokdad et al., 2003 and Yun et al., 2006)

| | <25 | 25–30 | 30+ |
|---|---|---|---|
| Diabetics | 2,545.9 | 2,607.5 | 9,517.6 |
| Nondiabetics | 74,977.1 | 55,776.3 | 49,580.6 |
| OR | Reference | 1.4 | 5.7 |



University at Buffalo The State University of New York

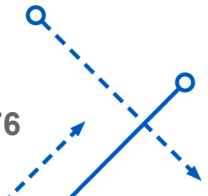
# MULTIDIMENSIONAL BIAS ANALYSIS

# Concept

- Conceptually, multidimensional bias analysis is no different than the three ‘simple’ types we have already discussed

- **There is one important distinction:**
  - Rather than assuming you know the exact value of the bias parameters, you explore several different options for the values of the bias parameters

- Repeat the simple bias analysis several times with different values of the bias parameters



# Example 1.

**TABLE 2.** *Classification of obesity using BMI >30 kg/m$^2$ and varying percent body fat cut-points to define obesity (n = 2,726)*

| | 35% Body fat | 38% Body fat | 40% Body fat |
|---|---|---|---|
| Prevalence of BMI >30 kg/m$^2$ | 66% | 45% | 32% |
| Sensitivity (95% CI) | 32.4% (29.5, 35.3) | 44.6% (40.9, 48.2) | 55.2% (50.9, 59.3) |
| Specificity (95% CI) | 99.3% (98.6, 99.8) | 97.1 (96.0, 98.1) | 94.6% (93.2, 95.9) |
| Positive predictive value (95% CI) | 98.8% (97.8, 99.8) | 92.5% (89.8, 95.1) | 82.9% (79.0, 86.8) |
| Negative predictive value (95% CI) | 43.6% (40.7, 46.4) | 68.5% (65.9, 71.0) | 81.5% (79.4, 83.6) |
| Percent correctly classified | 55.4% | 73.7% | 81.8% |
| Area under ROC curve | 0.658 | 0.708 | 0.749 |

| | Obese+ | Obese- |
|---|---|---|
| **Diabetes+** | 583 | 1437 |
| **Diabetes-** | 3853 | 1664 |

Crude RR= 2.1 (1.9, 2.3)
Crude OR= 2.3 (2.0, 2.5)

What is the corrected OR for 35% Body Fat?

What is the corrected OR for 38% Body Fat?

What is the corrected OR for 40% Body Fat?

This spreadsheet can be used to conduct a simple sensitivity analysis to correct for outcome misclassification. The example follows the example in chapter 6.

Reset Example | Clear Data

## Input Bias Parameters

| | |
|---|---|
| Se (obese+) | 0.32 |
| Se (obese-) | 0.32 |
| Sp (obese+) | 0.99 |
| Sp (obese-) | 0.99 |

| Variable Names | |
|---|---|
| Outcome | diab |
| Exposure | obese |

Error Check: No errors

## Instructions

Enter the bias parameters in the blue cells to the left and the observed data in the blue cells below. Cells in green give the results after correcting for outcome misclassification. Note: green cells are expected values and therefore do not have to be integers.

## Data (Enter obese-diab Data in Blue Cells)

Observed Data

| | obese + | obese - | Total |
|---|---|---|---|
| diab + | 583 a | 854 b | 1437 |
| diab - | 3853 c | 12788 d | 16641 |
| Total | 4436 m | 13642 n | |

| Observed | Measure (95% CI) |
|---|---|
| RR (obese-diab) | 2.1 (1.9 - 2.32) |
| OR (obese-diab) | 2.27 (2.03 - 2.53) |

## Corrected Data

Corrected Data

| | obese + | obese - | Total |
|---|---|---|---|
| diab + | 1737.5 A | 2314.8 B | 4052.3 |
| diab - | 2698.5 C | 11327.2 D | 14025.7 |
| Total | 4436 M | 13642 N | |

| Corrected | Measure |
|---|---|
| RR (obese-diab) | 2.31 |
| OR (obese-diab) | 3.15 |

## OUTCOME MISCLASSIFICATION

Chapter 6

This spreadsheet can be used to conduct a simple sensitivity analysis to correct for outcome misclassification. The example follows the example in chapter 6.

Reset Example | Clear Data

### Input Bias Parameters

| | |
|---|---|
| Se (obese+) | 0.45 |
| Se (obese-) | 0.45 |
| Sp (obese+) | 0.97 |
| Sp (obese-) | 0.97 |

**Variable Names**

| | |
|---|---|
| Outcome | diab |
| Exposure | obese |

Error Check: No errors

### Instructions

Enter the bias parameters in the blue cells to the left and the observed data in the blue cells below. Cells in green give the results after correcting for outcome misclassification. Note: green cells are expected values and therefore do not have to be integers.

### Data (Enter obese-diab Data in Blue Cells)

**Observed Data**

| | obese + | obese - | Total |
|---|---|---|---|
| diab + | 583 (a) | 854 (b) | 1437 |
| diab - | 3853 (c) | 12788 (d) | 16641 |
| Total | 4436 (m) | 13642 (n) | |

| Observed | Measure (95% CI) |
|---|---|
| RR (obese-diab) | 2.1 (1.9 - 2.32) |
| OR (obese-diab) | 2.27 (2.03 - 2.53) |

### Corrected Data

**Corrected Data**

| | obese + | obese - | Total |
|---|---|---|---|
| diab + | 1081.5 (A) | 1058.9 (B) | 2140.4 |
| diab - | 3354.5 (C) | 12583.1 (D) | 15937.6 |
| Total | 4436 (M) | 13642 (N) | |

| Corrected | Measure |
|---|---|
| RR (obese-diab) | 3.14 |
| OR (obese-diab) | 3.83 |

This spreadsheet can be used to conduct a simple sensitivity analysis to correct for outcome misclassification. The example follows the example in chapter 6.

Reset Example | Clear Data

## Input Bias Parameters

| | |
|---|---|
| Se (obese+) | 0.55 |
| Se (obese-) | 0.55 |
| Sp (obese+) | 0.95 |
| Sp (obese-) | 0.95 |

| Variable Names | |
|---|---|
| Outcome | diab |
| Exposure | obese |

Error Check: No errors

## Instructions

Enter the bias parameters in the blue cells to the left and the observed data in the blue cells below. Cells in green give the results after correcting for outcome misclassification. Note: green cells are expected values and therefore do not have to be integers.

## Data (Enter obese-diab Data in Blue Cells)

Observed Data

| | obese + | obese - | Total |
|---|---|---|---|
| diab + | 583 a | 854 b | 1437 |
| diab - | 3853 c | 12788 d | 16641 |
| Total | 4436 m | 13642 n | |

| Observed | Measure (95% CI) |
|---|---|
| RR (obese-diab) | 2.1 (1.9 - 2.32) |
| OR (obese-diab) | 2.27 (2.03 - 2.53) |

## Corrected Data

Corrected Data

| | obese + | obese - | Total |
|---|---|---|---|
| diab + | 722.4 A | 343.8 B | 1066.2 |
| diab - | 3713.6 C | 13298.2 D | 17011.8 |
| Total | 4436 M | 13642 N | |

| Corrected | Measure |
|---|---|
| RR (obese-diab) | 6.46 |
| OR (obese-diab) | 7.52 |

# Example 2.

Study examining the relationship between antidepressant use and cancer risk (Chien et al., 2006):

- Case control study of women aged 65-79 resident in Washington state
- Cases from SEER registry, controls from CMS
- In person interview to ascertain case status

- Conclusion: “No association between ever use of antidepressants and breast cancer risk (OR=1.2 ; 95% CI: 0.9, 1.6)”

| | Ever AD | Never AD |
|---|---|---|
| Cases | 118 | 832 |
| Controls | 103 | 884 |
| Crude OR | 1.21 (0.92, 1.61) | |
| Adjusted OR | 1.2 (0.9, 1.6) | |

# Could it be due to self-report bias?

- Used pharmacy records to validate self-report AD use:

**Cases**

$$\text{sensitivity} = \frac{24}{24+19} = 56\%$$

$$\text{specificity} = \frac{144}{144+2} = 99\%$$

**Controls**

$$\text{sensitivity} = \frac{18}{18+13} = 58\%$$

$$\text{specificity} = \frac{130}{130+4} = 97\%$$

| SE =57%<br>SP =98% | Observed | | Expected Truth | |
|---|---|---|---|---|
| | Ever AD | Never AD | Ever AD | Never AD |
| Cases | 118 | 832 | 183.3 | 766.7 |
| Controls | 103 | 884 | 154.2 | 832.8 |
| OR | 1.2 | | 1.3 | |

# Varying the bias parameters

| | Controls | | | | |
|---|---|---|---|---|---|
| **Cases** | **Sensitivity** | **0.6** | **0.5** | **0.6** | **0.5** |
| **Sensitivity** | **Specificity** | **1** | **1** | **0.95** | **0.95** |
| **0.6** | **1** | 1.24 | 0.99 | 1.11 | 0.86 |
| **0.5** | **1** | 1.57 | 1.25 | 1.41 | 1.09 |
| **0.6** | **0.95** | 1.57 | 1.28 | 1.42 | 1.8 |
| **0.5** | **0.95** | 1.98 | 1.62 | 1.8 | 1.44 |

In this example, under non-differential misclassification the corrected OR are further from the null than the uncorrected estimate reported by the authors (=1.2).

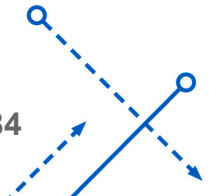
# Example 3.

Proportion

| | HBP | Normal BP | Total | Exposed | |
|---|---|---|---|---|---|
| Dead | 53 | 49 | 102 | 0.5196 | |
| Alive | 2125 | 3263 | 5388 | 0.3944 | |
| Total | 2178 | 3312 | 5490 | 0.3967 | |
| | Point estimate | | [95% Conf. | Interval] | |
| Odds ratio | 1.660879 | | 1.100383 | 2.510239 | (exact) |

You suspect selection bias, but have no way of estimating selection probabilities

Can take a ‘sensitivity analysis approach to it and vary the selection bias factor to see how the estimate changes



# Example 3.

| Selection Bias Factor | Corrected RR estimate | 95% Confidence Intervals |
|---|---|---|
| 0.3 | 5.54 | (3.7, 8.2) |
| 0.4 | 4.15 | (2.8, 6.2) |
| 0.5 | 3.32 | (2.2, 4.90 |
| 0.6 | 2.77 | (1.9, 4.1) |
| 0.7 | 2.37 | (1.6, 3.5) |
| 0.8 | 2.08 | (1.4, 3.1) |
| 0.9 | 1.85 | (1.2,2.7) |

Divide biased estimate by selection bias factor (and same for CIs) to get corrected estimate

```
episensi 53 49 2125 3263, st(cc) dsbfactor(c.(0.4))
```

# Problem with multidimensional bias analysis

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

Original article

Does selection bias explain the obesity paradox among individuals with cardiovascular disease?

Hailey R. Banack MA*, Jay S. Kaufman PhD

*Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, QC, Canada*

**Table 3**
Results of the sensitivity analysis for the controlled direct effect of obesity on mortality

| γ | Δ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
| −0.1 | −0.11 (−0.19, −0.03) | −0.10 (−0.18, −0.02) | −0.09 (−0.17, −0.01) | −0.08 (−.016, 0.003) | −0.07 (−0.15, 0.01) | −0.06 (−0.14, 0.02) | −0.051 (−0.13, 0.03) | −0.04 (−0.12, 0.12) | −0.03 (−0.11, 0.05) |
| −0.2 | −0.10 (−0.18, −0.02) | −0.08 (−0.16, 0.003) | −0.06 (−0.14, 0.02) | −0.04 (−0.12, 0.04) | −0.02 (−0.10, 0.06) | −0.001 (−0.08, 0.08) | 0.02 (−0.06, 0.10) | 0.04 (−0.04, 0.12) | 0.06 (−0.02, 0.14) |
| −0.3 | −0.09 (−0.17, −0.01) | −0.06 (−0.14, 0.02) | −0.03 (−0.11, 0.05) | −0.001 (−0.08, 0.08) | 0.03 (−0.05, 0.11) | 0.06 (−0.02, 0.14) | 0.09 (0.006, 0.17) | 0.12 (0.04, 0.20) | 0.15 (0.07, 0.23) |
| −0.4 | −0.08 (−0.16, 0.003) | −0.04 (−0.12, 0.04) | −0.08 (−0.08,0.08) | 0.04 (−0.04, 0.12) | 0.08 (−0.004, 0.08) | 0.12 (0.04, 0.20) | 0.16 (0.07, 0.24) | 0.20 (0.12, 0.28) | 0.24 (0.16, 0.32) |
| −0.5 | −0.07 (−0.15, 0.01) | −0.02 (−0.10, 0.06) | 0.03 (−0.05, 0.11) | 0.08 (−0.004, 0.16) | 0.13 (0.05, 0.21) | 0.18 (0.10, 0.26) | 0.23 (0.15, 0.31) | 0.28 (0.20, 0.36) | 0.33 (0.25, 0.41) |
| −0.6 | −0.06 (−0.14, 0.02) | −0.001 (−0.08, 0.08) | 0.06 (−0.02, 0.14) | 0.12 (0.04,0.20) | 0.18 (0.10, 0.26) | 0.24 (0.16, 0.32) | 0.30 (0.22, 0.38) | 0.36 (0.28, 0.44) | 0.42 (0.34, 0.50) |
| −0.7 | −0.05 (−0.13, 0.03) | 0.02 (−0.06,0.10) | 0.09 (−0.01, 0.17) | 0.16 (0.08, 0.24) | 0.23 (0.15, 0.31) | 0.30 (0.22, 0.38) | 0.37 (0.29, 0.45) | 0.44 (0.36, 0.52) | 0.51 (0.43, 0.59) |
| −0.8 | −0.04 (−0.12, 0.04) | 0.04 (−0.04, 0.12) | 0.12 (0.04, 0.20) | 0.20 (0.12, 0.28) | 0.28 (0.20, 0.36) | 0.36 (0.28, 0.44) | 0.44 (0.36, 0.52) | 0.52 (0.44, 0.60) | 0.60 (0.52, 0.68) |
| −0.9 | −0.03 (−0.11,0.05) | 0.06 (−0.20, 0.14) | 0.15 (0.07, 0.23) | 0.24 (0.16, 0.32) | 0.33 (0.25,0.41) | 0.42 (0.34, 0.50) | 0.51 (0.43, 0.60) | 0.60 (0.52, 0.68) | 0.69 (0.61, 0.77) |

Shading demonstrates direction of effect. The white background indicates the parameters for which obesity appears protective, light gray shading demonstrates a null effect, and dark gray shading demonstrates obesity is harmful.

1. It is really hard to present results
   - Lots of space in text and tables may make editors ☹

2. Don't get any understanding of which combination of bias parameters is the 'true' combination

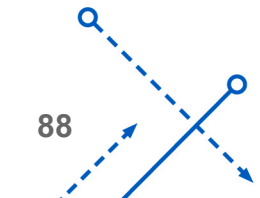
University at Buffalo The State University of New York

# PROBABILISTIC BIAS ANALYSIS

# Concept

- Probabilistic bias analysis is an extension of simple bias analysis in which the  analyst assigns probability distributions to the bias parameters, rather than single values

- Repeatedly sampling from the probability distributions using Monte Carlo Sampling techniques and correcting for the bias, the result is a frequency distribution of revised estimates of association



# Probabilistic bias analysis

**1. Specify a probability distribution for the bias parameters**

**2. Use Monte Carlo sampling techniques to select bias parameter values from the probability distributions**

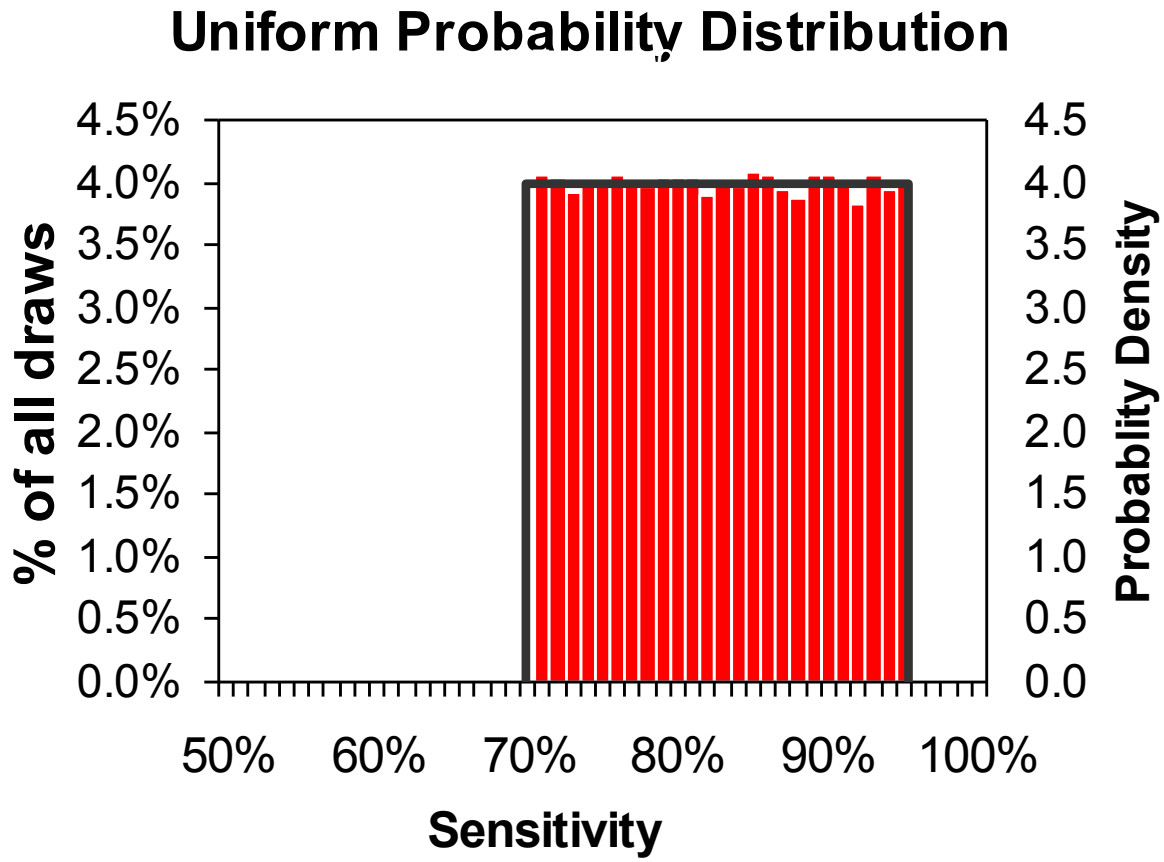
**3. Use the sampled bias parameter values to correct the 'naive' effect estimate**

**4. Repeat steps 2 and 3 (many times)**

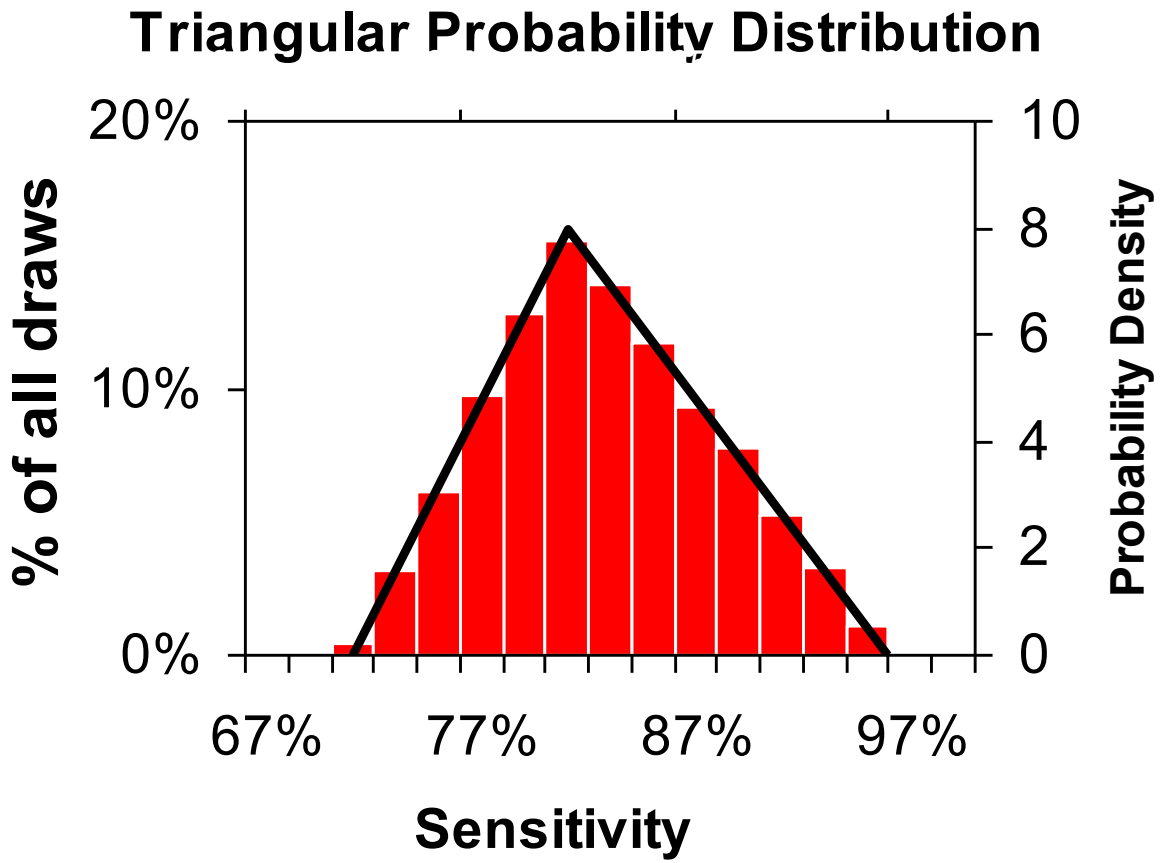
**5. Generate a frequency distribution of corrected effect estimates**

- $50^{th}$ percentile of this distribution = bias-corrected risk ratio
- 2.5 and 97.5 percentiles = upper and lower limits of 95% CI

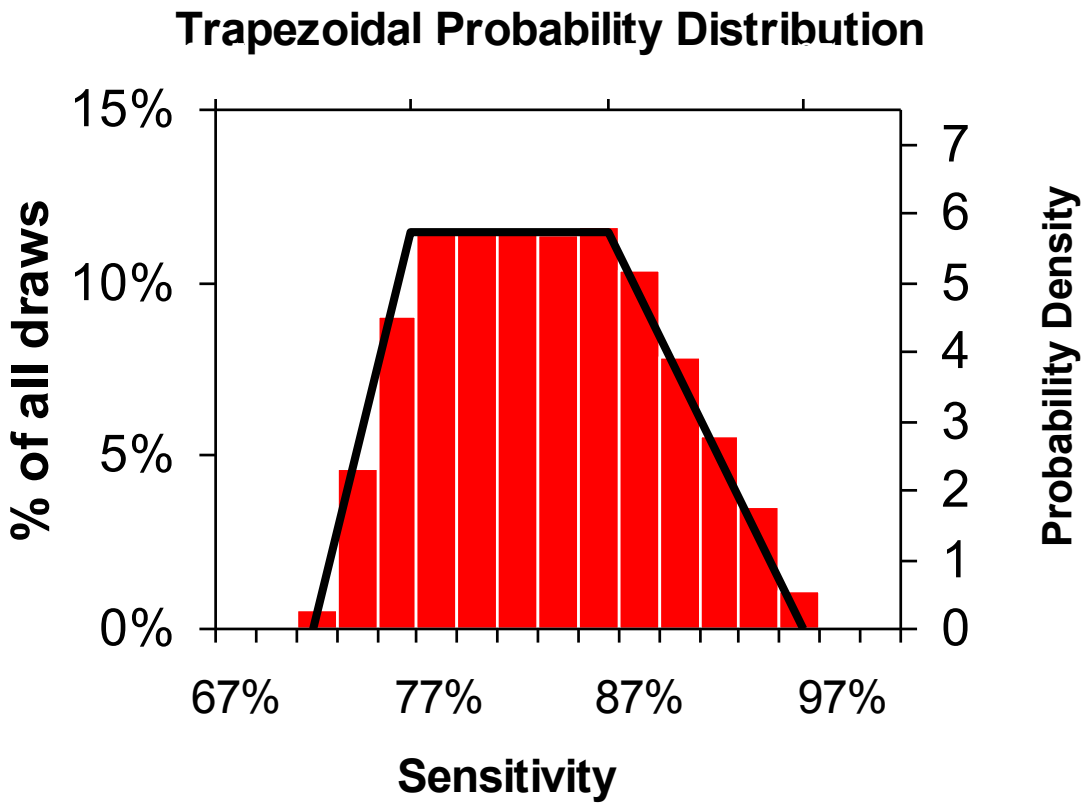
# Probability Distributions: Uniform



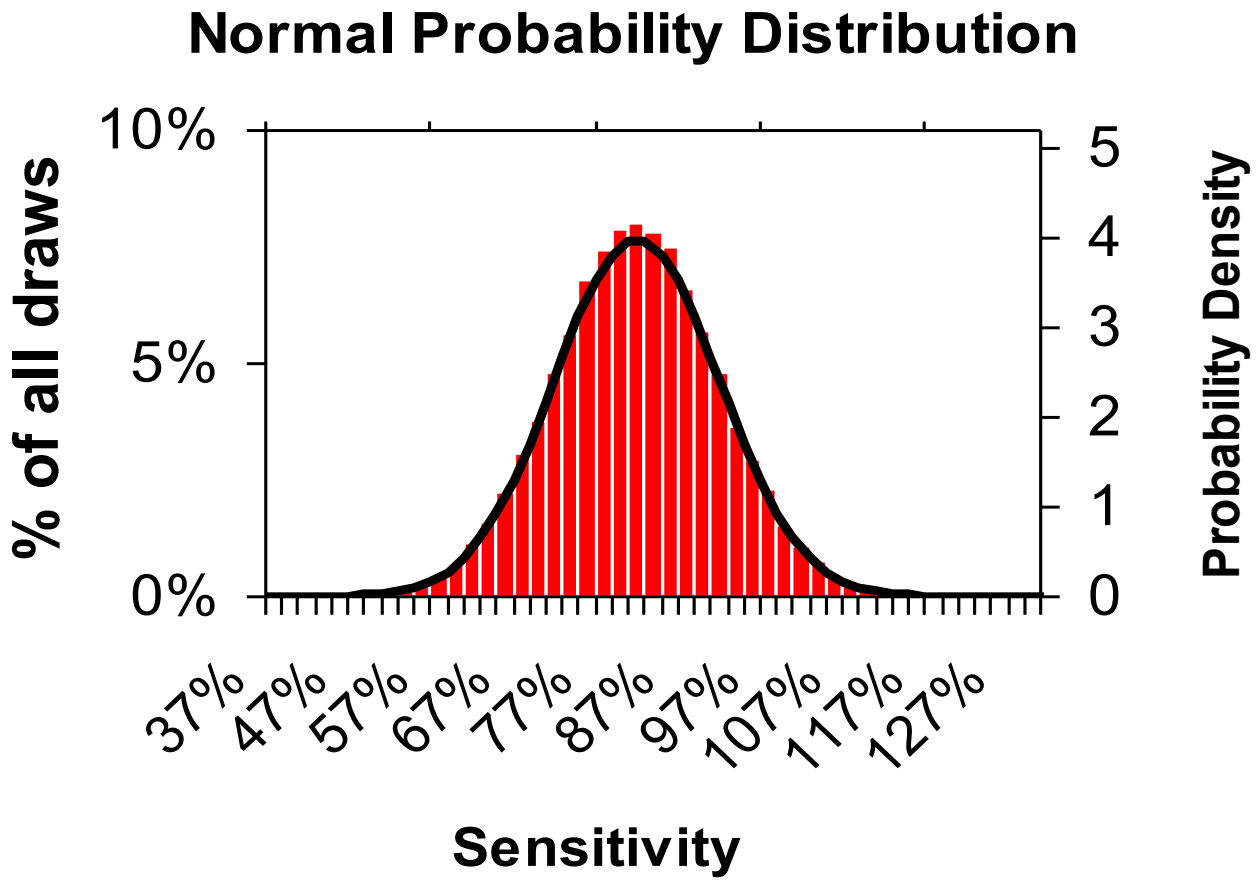
# Probability Distributions: Triangular



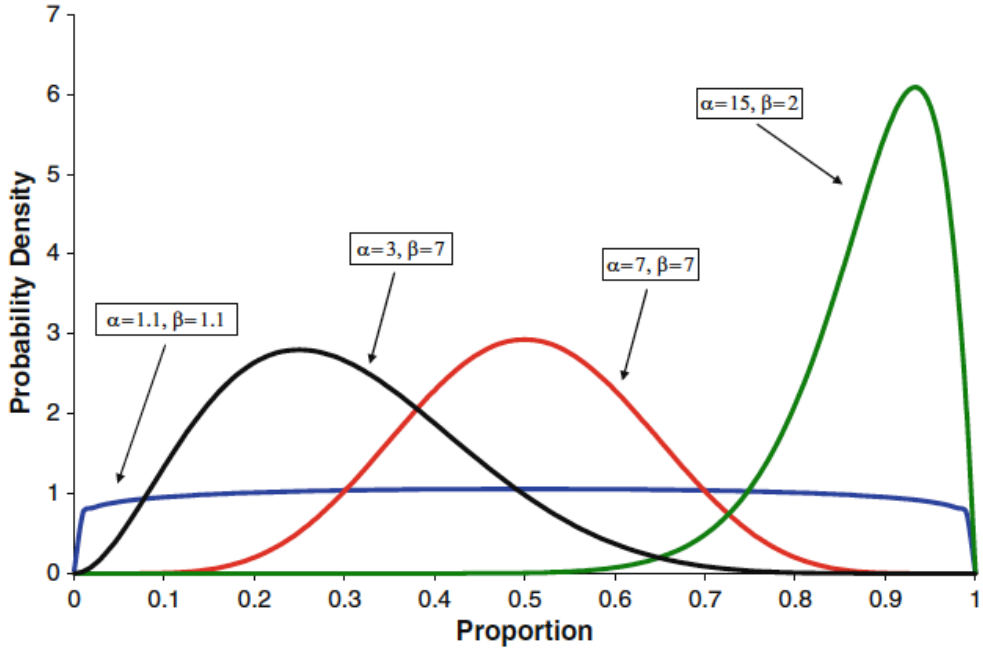
# Probability Distributions: Trapezoidal



# Probability Distributions: Normal



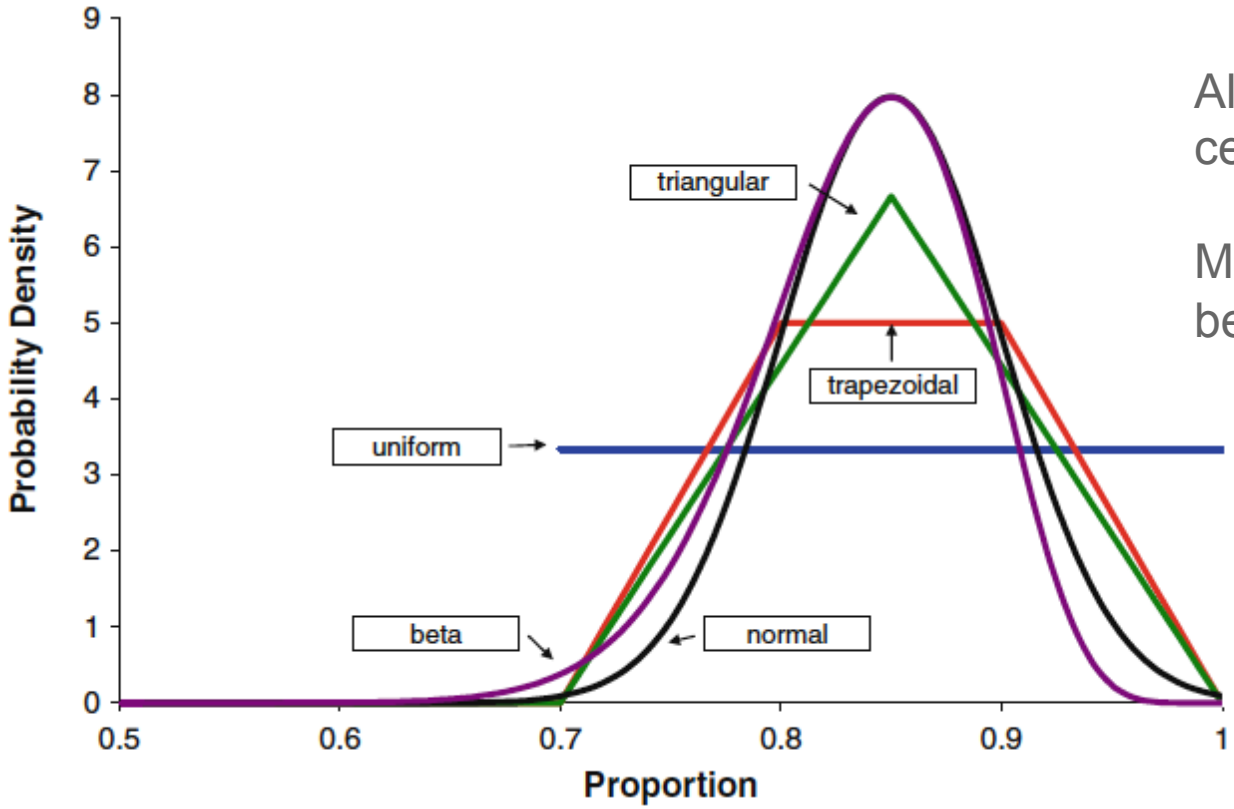
# Probability Distributions: Beta




**Beta**

- Works well for proportions and validation data
- For sensitivity, among true positives, set
  Alpha = N test positives + 1
  Beta = N test negatives + 1

# Comparing distributions




All of the distributions here are centered around 0.85

Majority of the density lies between 0.7 and 0.1

Choice of distribution not likely to have a major effect on the results of the bias analysis

# Software options

**Stata:**

- rbeta
- runiform
- episens has several probability density functions built-in

**SAS:**

- standard uniform: ranuni(seed)
- quantile('uniform',u,{l,r})
- rannor(seed) ← standard normal
- quantile('normal',u, μ,σ)
- quantile('beta',u,α,β,l,r)

# -episens- for probabilistic bias analysis

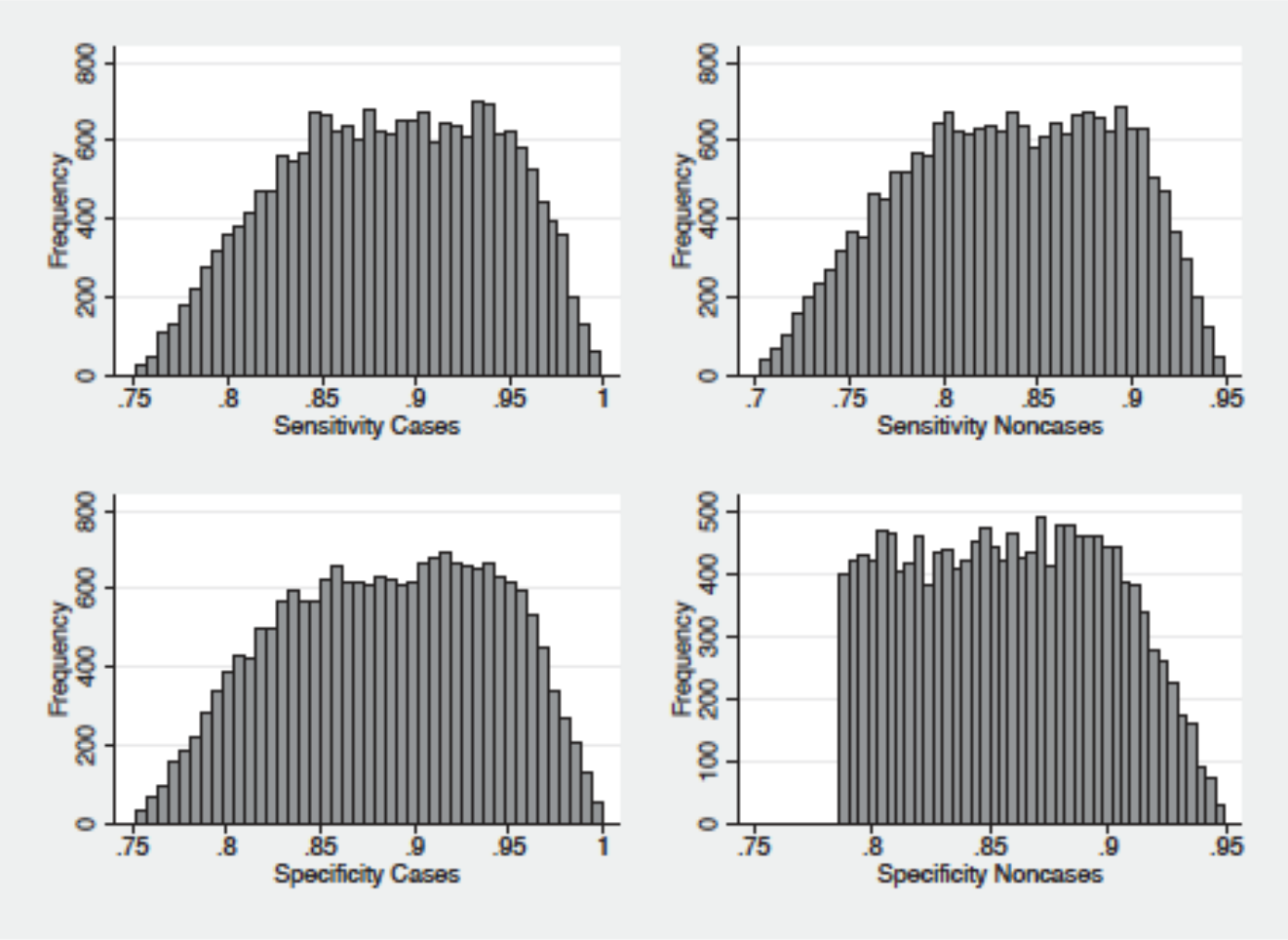
| | Exposed | Unexposed | Total | Proportion Exposed | |
|---|---|---|---|---|---|
| Cases | 45 | 94 | 139 | 0.3237 | |
| Controls | 257 | 945 | 1202 | 0.2138 | |
| Total | 302 | 1039 | 1341 | 0.2252 | |
| | Point estimate | | [95% Conf. | Interval] | |
| Odds ratio | 1.760286 | | 1.202457 | 2.576898 | (Woolf) |
| Attr. frac. ex. | .4319106 | | .1683693 | .6119365 | (Woolf) |
| Attr. frac. pop | .1398272 | | | | |

chi2(1) = 8.63 Pr>chi2 = 0.0033

Case Control study: Exposure to resins & lung cancer cases (Greenland et al., 1994)

```
. episensi 45 94 257 945, st(cc) reps(20000) nodots dseca(trap(.75 .85 .95 1))
> dspca(trap(.75 .85 .95 1)) dsenc(trap(.7 .8 .9 .95))
> dspnc(trap(.7 .8 .9 .95)) corrsens(.8) corrspec(.8) seed(123) grprior
Se|Cases    : Trapezoidal(.75,.85,.95,1)
Sp|Cases    : Trapezoidal(.75,.85,.95,1)
Se|No-Cases: Trapezoidal(.7,.8,.9,.95)
Sp|No-Cases: Trapezoidal(.7,.8,.9,.95)
Corr Se|Cases and Se|No-Cases : .8
Corr Sp|Cases and Sp|No-Cases : .8
Probabilistic sensitivity analysis for misclassification of the exposure
                                    Percentiles          Ratio
                              2.5       50        97.5      97.5/2.5
                              ----------------------------------------
Conventional                  1.20      1.76      2.58      2.14
Systematic error              1.81      3.48      48.19     26.57
Systematic and random error   1.61      3.60      48.92     30.47
```

# Exposure misclassification histograms




Histograms of 20,000 draws from trapezoidal prior distributions for the sensitivity and specificity among cases and non-cases.

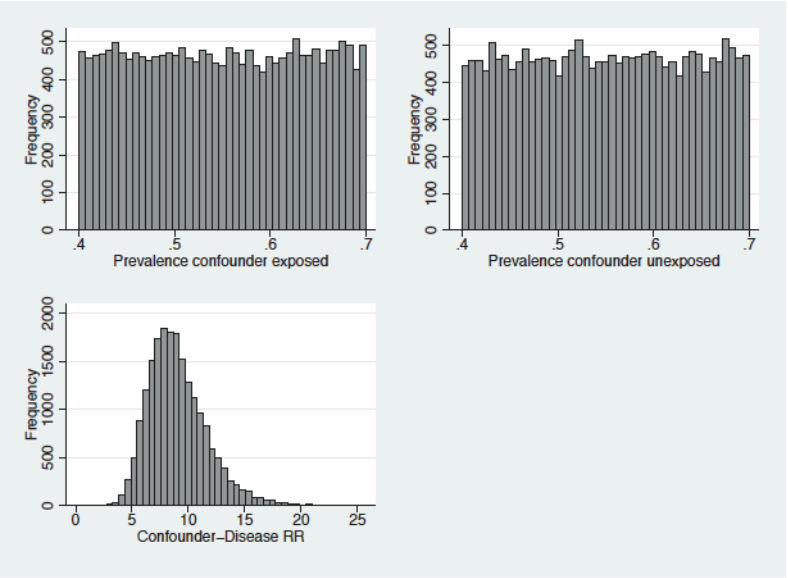
# Unmeasured confounder (smoking)

```
. episensi 45 94 257 945, st(cc) reps(20000) nodots dpexp(uni(.4 .7))
> dpunexp(uni(.4 .7)) drrcd(log-n(2.159 .280)) seed(123)
> grarrtot grprior

Pr(c=1|e=1): Uniform(.4,.7)
Pr(c=1|e=0): Uniform(.4,.7)
RR_cd      : Log-Normal(2.16,0.28)

Probabilistic sensitivity analysis for unmeasured confounding

                                  Percentiles          Ratio
                          2.5         50         97.5      97.5/2.5
                          ----------------------------------------
Conventional              1.20        1.76       2.58      2.14
Systematic error          1.25        1.76       2.49      2.00
Systematic and random error 1.05      1.76       2.96      2.83
```



# Selection Bias example

## Association of Weight Status With Mortality in Adults With Incident Diabetes

Mercedes R. Carnethon, PhD
Peter John D. De Chavez, MS
Mary L. Biggs, PhD
Cora E. Lewis, MD
James S. Pankow, PhD
Alain G. Bertoni, MD, MS
Sherita H. Golden, MD, MS
Kiang Liu, PhD
Kenneth J. Mukamal, MD, MPH
Brenda Campbell-Jenkins, PhD
Alan R. Dyer, PhD

**Objective** To test the association of weight status with mortality in adults with new-onset diabetes in order to minimize the influence of diabetes duration and voluntary weight loss on mortality.

**Design, Setting, and Participants** Pooled analysis of 5 longitudinal cohort studies: Atherosclerosis Risk in Communities study, 1990-2006; Cardiovascular Health Study, 1992-2008; Coronary Artery Risk Development in Young Adults, 1987-2011; Framingham Offspring Study, 1979-2007; and Multi-Ethnic Study of Atherosclerosis, 2002-2011. A total of 2625 participants with incident diabetes contributed 27 125 person-years of follow-up. Included were men and women (age >40 years) who developed incident diabetes based on fasting glucose 126 mg/dL or greater or newly initiated diabetes medication and who had concurrent measurements of BMI. Participants were classified as normal weight if their BMI was 18.5 to 24.99 or overweight/obese if BMI was 25 or greater.

**Conclusion** Adults who were normal weight at the time of incident diabetes had higher mortality than adults who are overweight or obese.

*JAMA. 2012;308(6):581-590* www.jama.com

- Analytic cohort comprised of 2625 men and women with incident diabetes (based on fasting glucose or newly initiated medication for diabetes)

**Table 2.** Association Between Weight Status at the Time of Diabetes Incidence and Mortality[a]

| | Total Mortality | | Cardiovascular Mortality | |
|---|---|---|---|---|
| | Overweight/Obese | Normal Weight | Overweight/Obese | Normal Weight |
| | Full Sample | | | |
| No. of participants | 2332 | 293 | 2086 | 265 |
| No. of events | 371 | 78 | 149 | 24 |
| Event rate (per 10 000 person-years) | 152.1 | 284.8 | 67.8 | 99.8 |
| Pooled analysis[b] | | | | |
| Unadjusted HR (95% CI) | 1 [Reference] | 1.70 (1.33-2.18) | 1 [Reference] | 1.31 (0.85-2.02) |
| Multivariable model 1[c] | 1 [Reference] | 1.49 (1.15-1.93) | 1 [Reference] | 1.04 (0.65-1.66) |
| Multivariable model 2[d] | 1 [Reference] | 2.08 (1.52-2.85) | 1 [Reference] | 1.52 (0.89-2.58) |

| | Normal weight | Overweight/obese |
|---|---|---|
| Dead | 78 | 371 |
| Alive | 215 | 1961 |

Carnethon paper

| | Normal weight | Overweight/obese |
|---|---|---|
| Dead | 1835 | 2735 |
| Alive | 5705 | 7812 |

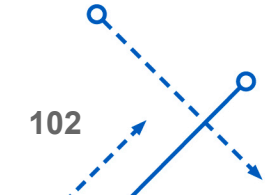
NHANES III
Complete cohort

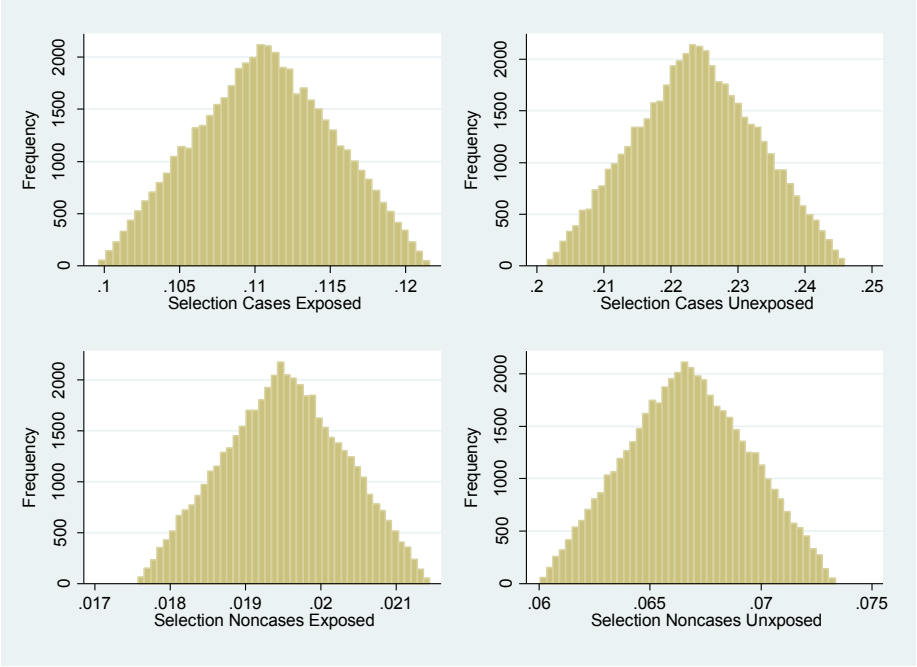
| | Normal weight | Overweight/obese |
|---|---|---|
| Dead | 203 | 612 |
| Alive | 99 | 521 |

NHANES III
Diabetics

| | Normal weight | Overweight/obese |
|---|---|---|
| Dead | 0.11 | 0.22 |
| Alive | 0.02 | 0.07 |

Sampling Fractions






| | HR Estimate from article | Crude OR Estimate | Selection bias corrected OR estimate |
|---|---|---|---|
| Normal weight | 1.70 (1.33, 2.18) | 1.92 (1.45, 2.54) | 1.13 (0.82, 1.57) |

*Overweight-obese (BMI>25) used as referent group

University at Buffalo The State University of New York

# ADVANCED TOPICS AND EXTENSIONS

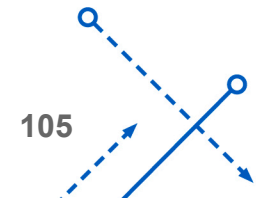
# Record-level vs. Summary Data

-Rather than using summary data as we have been

-If you have access to the dataset, can use record level corrections and then run analyses

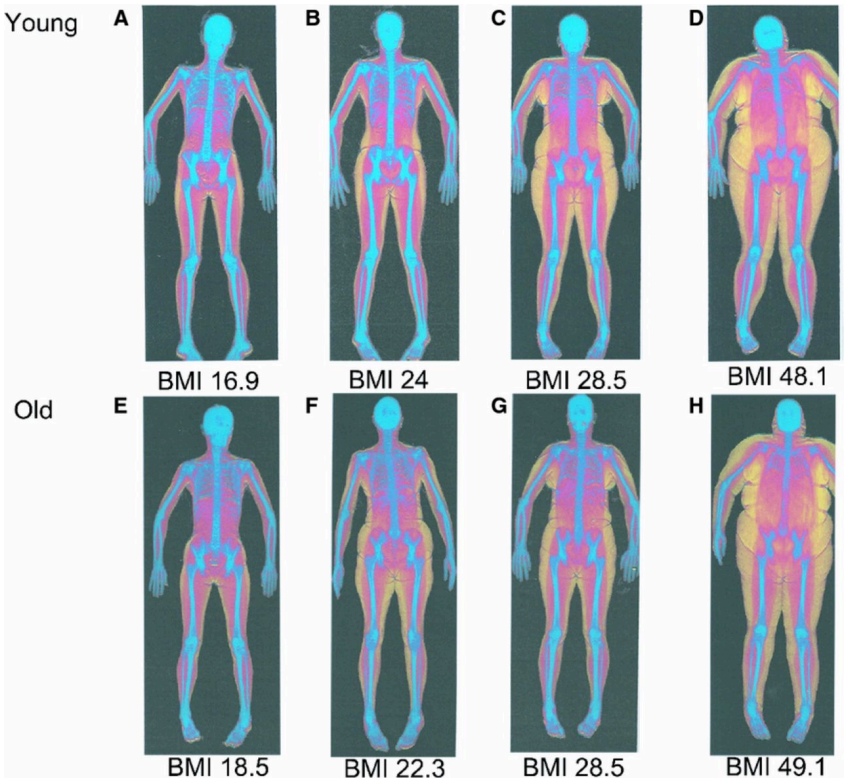
-Shift from correcting biased effect estimates to ‘correcting’ data before analysis

-Really, really, really computationally intensive



# Example: Age-related BMI misclassification

Does the effect of obesity on mortality change when accounting for BMI-related misclassification?




**As women age:**

- Increased body fat
- Changes in the distribution of adipose tissue
- Decreased lean body mass
- Loss of height
- Loss of bone density

# Calculating Bias Parameters

Use sensitivity and specificity from validation study to calculate PPV and NPV

Calculate the number of expected number of exposed and unexposed individuals at each level of the outcome

Observed

| | E=1 | E=0 | Total |
|---|---|---|---|
| D=1 | a | b | a+b |
| D=0 | c | d | c+d |
| Total | a+c | b+d | n |

Expected

| | E=1 | E=0 | Total |
|---|---|---|---|
| D=1 | A | B | A+B |
| D=0 | C | D | C+D |
| Total | A+C | B+D | N |

**A = [a-(1-sensitivity) * N] / [sensitivity-(1-specificity)**
**B = N-A**
**C = [c-(1- sensitivity) * N] / [sensitivity -(1- specificity)]**
**D = N-C**

This information is then used to calculate the number of expected true positives (TP), true negatives (TN), false positive (FP) and false negatives (FN) across levels of the outcome

| D=1 | D=0 |
|---|---|
| $TP_{D=1}$ = Sensitivity * A | $TP_{D=0}$ = Sensitivity * C |
| $TN_{D=1}$ = Specificity * B | $TN_{D=0}$ = Specificity * D |
| $FP_{D=1}$ = (1-sensitivity) * B | $FP_{D=0}$ = (1-sensitivity) * D |
| $FN_{D=1}$ = (1-specificity) * A | $FN_{D=0}$ = (1-specificity) * C |

(Fox et al., 2005)

And finally, TP, TN, FP, and FN are used to calculate the PPV and NPV values for D=1 and D=0:

$$PPV_{D=1} = TP_{D=1} / (TP_{D=1} + FP_{D=1})$$

$$NPV_{D=1} = TN_{D=1} / (TN_{D=1} + FN_{D=1})$$

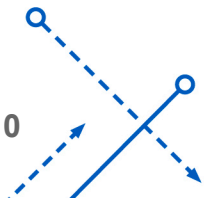
$$PPV_{D=0} = TP_{D=0} / (TP_{D=0} + FP_{D=0})$$

$$NPV_{D=0} = TN_{D=0} / (TN_{D=0} + FN_{D=0})$$

# Record-level correction

## Conduct Bernoulli trials:

- For each record in the dataset, a random number is chosen from a binomial distribution
- If the number chosen is **greater** than the corresponding predictive value, the participant's exposure status is reclassified
  - For women originally classified as obese by BMI, PPV is the probability of correct classification
  - For women originally classified as non-obese by BMI, NPV is the probability of correct classification
- Reclassification process repeated for each record (row) in the dataset to create a new, corrected dataset



# Record-level correction

| ID | Obese | PPV | NPV | New_Obese |
|---|---|---|---|---|
| 1 | 0 | 0.96 | **0.58** | 1 |
| 2 | 0 | 0.97 | **0.48** | 0 |
| 3 | 1 | **0.94** | 0.61 | 1 |
| 4 | 1 | **0.94** | 0.61 | 0 |
| 5 | 1 | **0.98** | 0.49 | 1 |
| 6 | 1 | **0.96** | 0.63 | 1 |
| 7 | 0 | 0.97 | **0.48** | 1 |

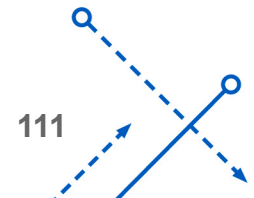
Number drawn greater than 0.58

Number drawn less than 0.48

Number drawn greater than 0.94

Number drawn less than 0.96

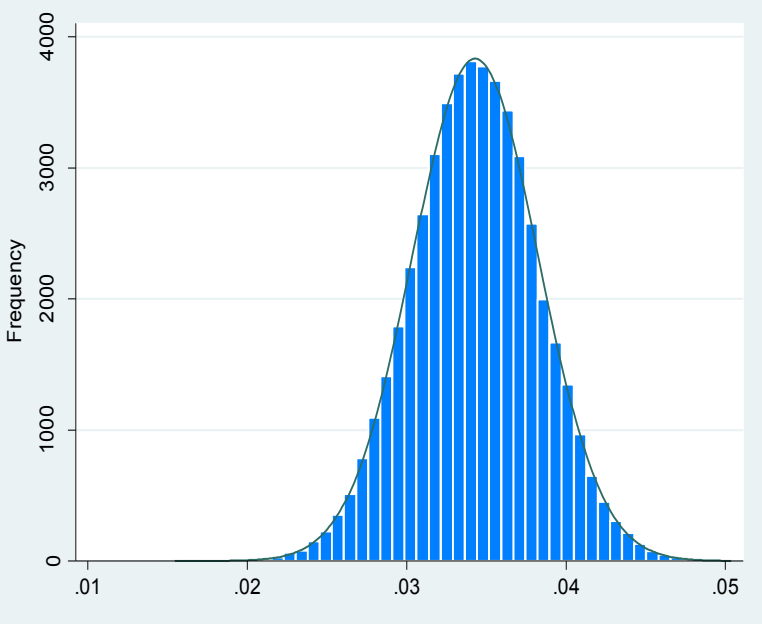
Go through this process for every record in the dataset to create a new obesity exposure value for each individual



# Repeat… many times

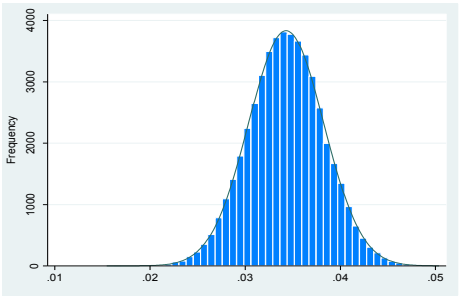
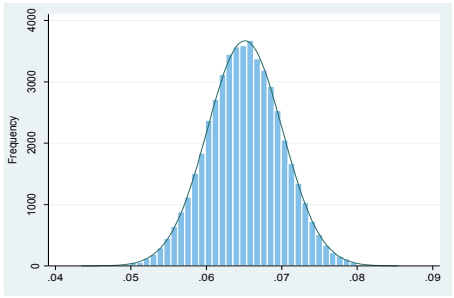
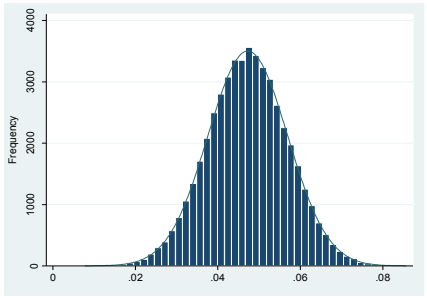
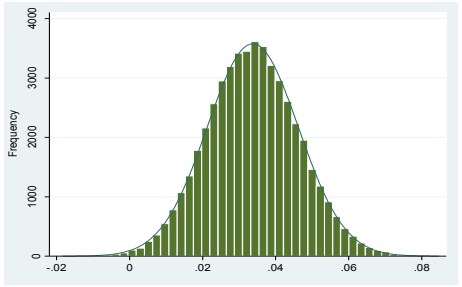
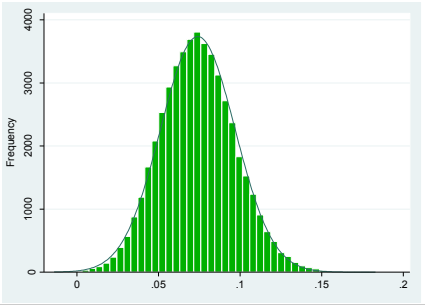
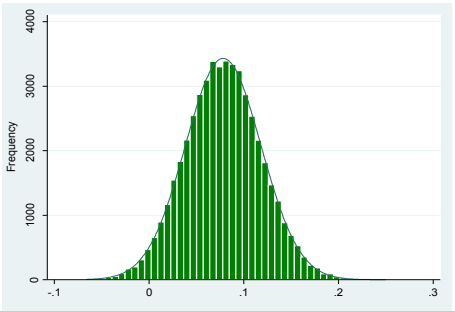
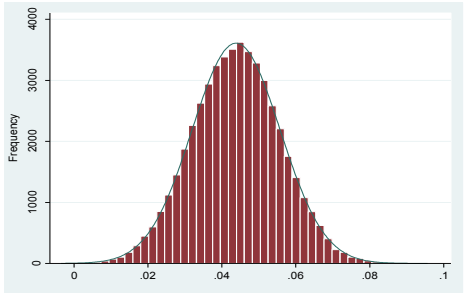
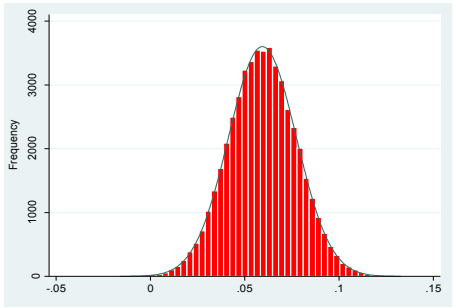
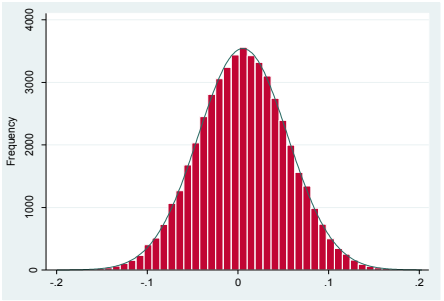
Repeat entire process 50,000 times to create a distribution of risk differences

Bias corrected effect estimate is the 50th percentile of the distribution for each race/ethnicity strata




2.5th and 97.5th percentiles of the distribution correspond to upper and lower 95% confidence intervals

# Results

| | Age 50-59 | Age 60-69 | Age 70-79 |
|---|---|---|---|
| **NH-White** |  |  |  |
| **NH-Black** |  |  |  |
| **Hispanic** |  |  |  |

# Bias Analysis Results

| | Uncorrected risk difference* (95% SI) | Bias-corrected risk difference* (95% SI) |
|---|---|---|
| **Non-Hispanic White** | | |
| **50-59** | 0.017 (0.010, 0.023) | 0.034 (0.026, 0.042) |
| **60-69** | 0.027 (0.019, 0.034) | 0.065 (0.055, 0.075) |
| **70-79** | 0.029 (0.014, 0.044) | 0.047 (0.028, 0.066) |
| **Non-Hispanic Black** | | |
| **50-59** | 0.015 (-0.002, 0.033) | 0.034 (0.010, 0.058) |
| **60-69** | 0.006 (-0.017, 0.029) | 0.074 (0.031, 0.12) |
| **70-79** | 0.027 (-0.025, 0.080) | 0.077 (0.001, 0.16) |
| **Hispanic** | | |
| **50-59** | 0.022 (0.002, 0.041) | 0.044 (0.02, 0.066) |
| **60-69** | 0.043 (0.014, 0.071) | 0.059 (0.024, 0.094) |
| **70-79** | -0.009 (-0.09, 0.070) | 0.005 (-0.092, 0.10) |

*Adjusted for smoking status, hormone therapy use, physical activity, years since menopause, education, income, employment status, marital status, and alcohol intake

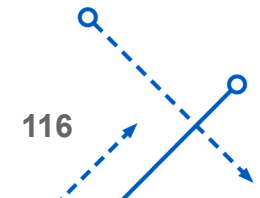
University at Buffalo The State University of New York

# WRAP UP & CONCLUSIONS

# Review

What we have covered today:

- Key concepts about bias and different types of bias
- Deterministic and probabilistic bias analysis techniques
  - Selection bias
  - Misclassification
  - Unmeasured confounding
- Hand calculations, Stata commands, Excel worksheets
- Didn't cover Bayesian bias analysis, multiple bias analysis, more complex scenarios (e.g., polytomous unmeasured confounders, EMM with unmeasured confounding)



Int. J. Epidemiol. Advance Access published July 30, 2014

*International Journal of Epidemiology*, 2014, 1–17
doi: 10.1093/ije/dyu149
Original article

Original article

# Good practices for quantitative bias analysis

**Timothy L Lash,[1]* Matthew P Fox,[2] Richard F MacLehose,[3] George Maldonado,[4] Lawrence C McCandless[5] and Sander Greenland[6]**

[1]Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA, USA, [2]Department of Epidemiology and Center for Global Health & Development, Boston University School of Public Health, Boston, MA, USA, [3]Division of Epidemiology and Community Health and [4]Division of Environmental Health Sciences, University of Minnesota School of Public Health, Minneapolis, MN, USA, [5]Faculty of Health Sciences, Simon Fraser University, Burnaby, BC, Canada and [6]Department of Epidemiology and Department of Statistics, University of California Los Angeles, Los Angeles, CA, USA

*Corresponding author. Department of Epidemiology, Rollins School of Public Health, Emory University, 1518-002-3BB, 1518 Clifton Rd NE, Atlanta, GA 30322, USA. E-mail: tlash@emory.edu

Accepted 7 July 2014

# Thank you!!

Jay Kaufman

Tim Lash

Matt Fox

Jean Wactawski-Wende

Maria Glymour

June Chan

UCSF



Email: hrbanack@buffalo.edu