

# ASHG EDUCATIONAL SESSION

## Observational Study Designs

Moyses Szklo, MD, MPH, DrPH

The Johns Hopkins Bloomberg School of Public Health

**NOTHING TO DISCLOSE**

# Observational Study Designs

- By definition, an observational study is one in which the investigator does not control “assignment” of the potential risk factor of interest (e.g., smoking, cytomegalovirus)
- Good company: Geology, Astrophysics, Ecology, etc.

# Observational Study Designs

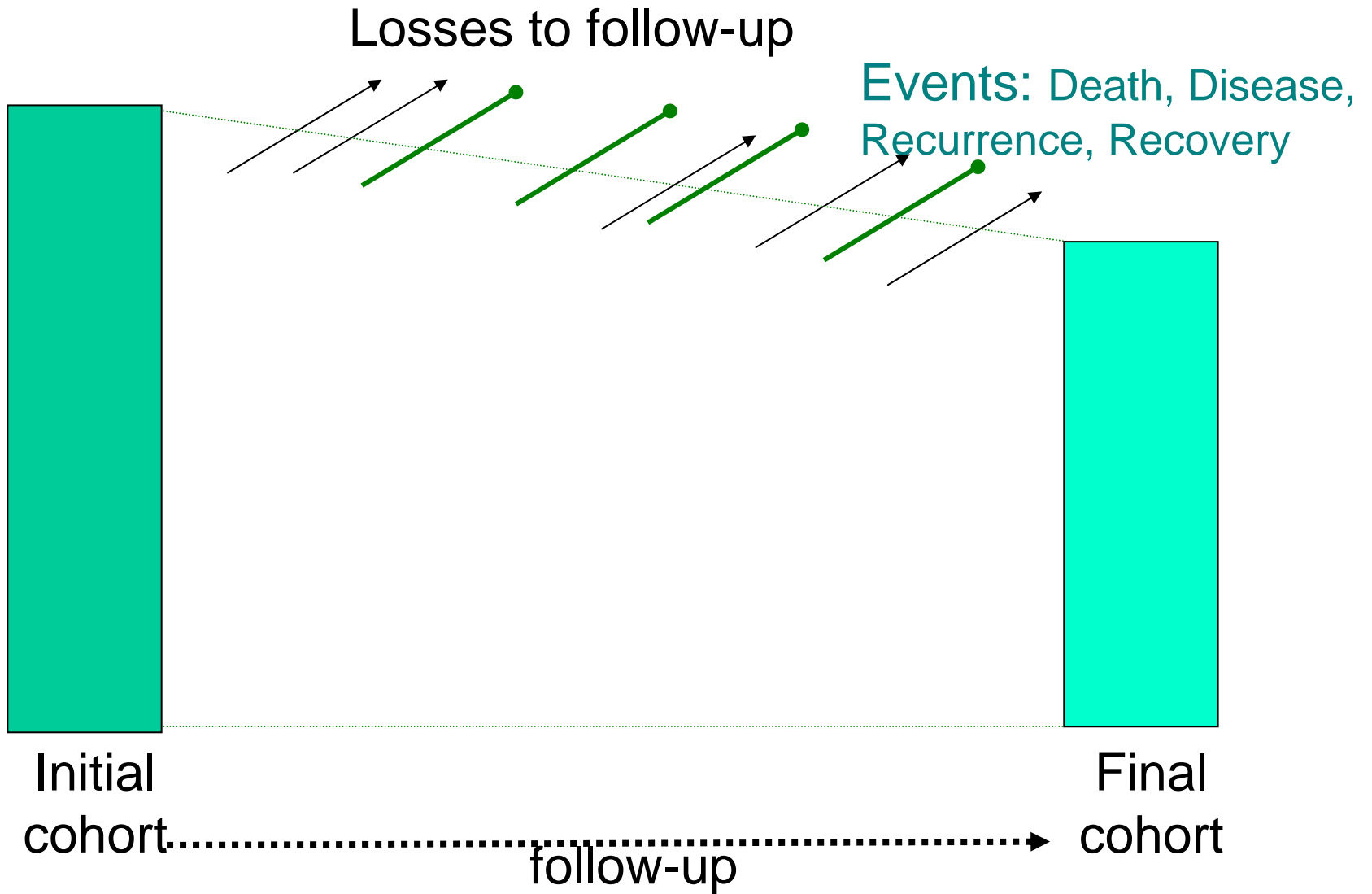
- Cohort

- Case-control

  - Traditional (case-based)

  - Case-cohort

# Cohort study



# Basic Design of a **Prospective (Cohort) Study** (Observational)

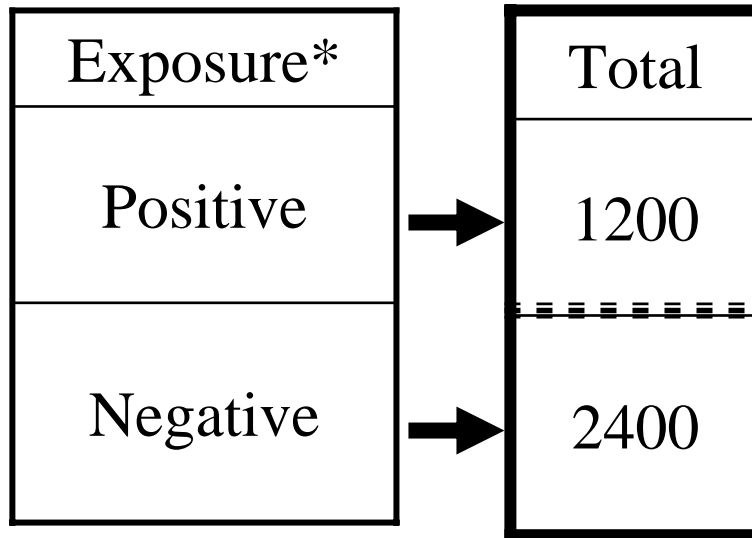
*First*, classify cohort by presence of exposure to the suspected risk factor:

Exposure*
Positive
Negative

(\*Example: smoking during pregnancy)

# Basic Design of a **Prospective (Cohort) Study** (Observational)

*First*, classify cohort by presence of exposure to the suspected risk factor:

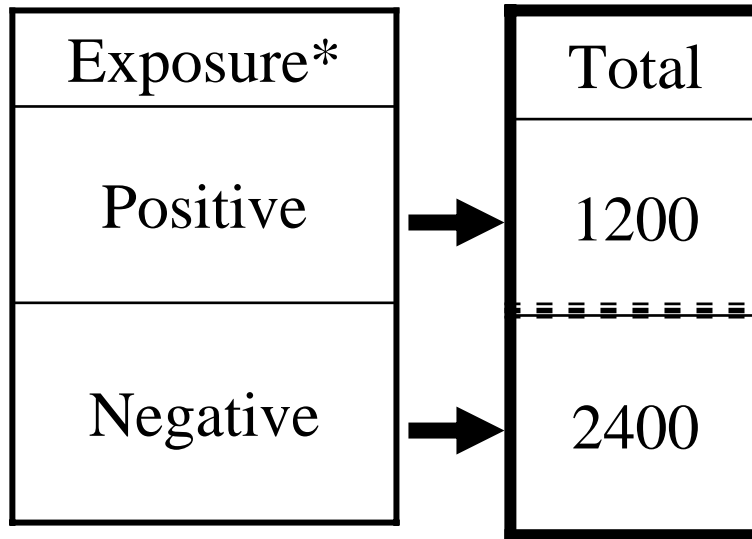


(\*Example: smoking during pregnancy)

# Basic Design of a **Prospective (Cohort) Study** (Observational)

*First*, classify cohort by presence of exposure to the suspected risk factor:

*Then*, follow up subjects to see who develops event (e.g., congenital malformation in offspring)



(\*Example: smoking during pregnancy)

# Basic Design of a **Prospective (Cohort) Study** (Observational)

*First*, classify cohort by presence of exposure to the suspected risk factor:

*Then*, follow up subjects to see who develops event (e.g., congenital malformation in offspring)

Exposure*	Total	Event	Non-event
Positive	1200	60	1140
Negative	2400	24	2376

(\*Example: smoking during pregnancy)

Incidence of event (e.g., congenital malformation):

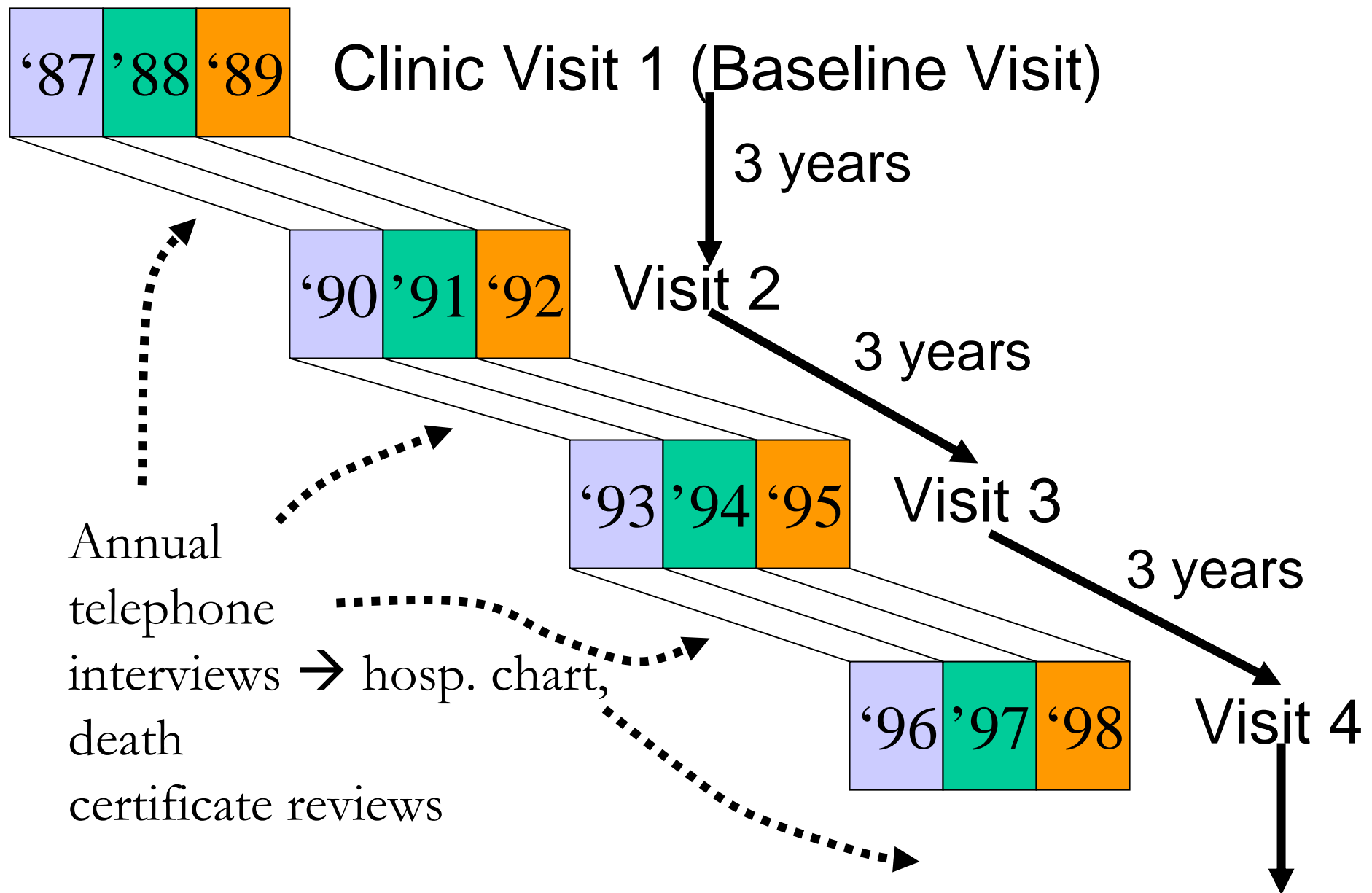
smokers:  $60/1200 = 5\%$   
non-smokers:  $24/2400 = 1\%$  } Relative Risk =  $5\% \div 1\% = 5.0$



## Atherosclerosis Risk in Communities (ARIC) Study

- Cohort (prospective) concurrent study to examine risk factors for subclinical and clinical atherosclerotic diseases
- Approximately 16,000 persons aged 45-64 yrs at baseline (1987-89)
- Multi-center: Jackson (all African-American), Forsyth County, NC (about 15% African-American), Minneapolis (mostly white) and Washington County, MD (mostly white)
- Follow-up approaches: Periodic visits to ARIC clinic; Annual telephone interviews → hospital chart and death certificate reviews

# Design of the ARIC Study



## Age-, Field Center- and Race-Adjusted Average Annual Coronary Heart Disease (CHD) Incidence Rates/1000, ARIC Cohort Study

Risk Factor	Women		Men
	Rate		Rate
Diabetes		<b>Difference in CHD risk between women and men decreases substantially when diabetes is present</b>	
Yes	9.2		13.8
No	1.8		6.4
Smoking		<b>CHD risk of former smokers is similar to that of never smokers</b>	
Current	5.3		11.5
Former	1.6		5.8
Never	1.3		4.7

First and often best way to analyze data (George Comstock): Before carrying out complex modeling, look at the data and think about what you are seeing!

# Measuring an Association Between a Suspected Risk Factor and a Disease

## Age-, Field Center- and Race-Adjusted Average Annual Coronary Heart Disease (CHD) Incidence Rates/1000, ARIC Cohort Study

Risk Factor	Women			Men		
	Rate	<b>RR</b>	<b>AR<sub>exp</sub>/1000</b>	Rate	<b>RR</b>	<b>AR<sub>exp</sub>/1000</b>
Diabetes						
Yes	9.2	<b>5.1</b>	<b>7.4</b>	13.8	<b>2.2</b>	<b>7.4</b>
No	1.8	<b>1.0</b>	<b>Ref.</b>	6.4	<b>1.0</b>	<b>Ref.</b>
Smoking						
Current	5.3	<b>4.1</b>	<b>4.0</b>	11.5	<b>2.4</b>	<b>6.8</b>
Former	1.6	<b>1.2</b>	<b>0.3</b>	5.8	<b>1.2</b>	<b>1.1</b>
Never	1.3	<b>1.0</b>	<b>Ref.</b>	4.7	<b>1.0</b>	<b>Ref.</b>

$$\text{Relative Risk} = \text{Incidence}_{\text{exp}} \div \text{Incidence}_{\text{unexp}}$$

**RR > 1.0 → Factor may be a risk factor**

**RR < 1.0 → Factor may be protective**

**RR = 1.0 → No association**

# Observational Study Designs

- Cohort

- Case-control

- Traditional (case-based)

- Case-cohort

# Traditional Case-Control Study

**First**, select cases with the disease of interest and disease-free controls:

# Traditional Case-Control Study

**First**, select cases with the disease of interest and disease-free controls:

Cases	Controls
-------	----------

# Traditional Case-Control Study

**Then**, ascertain past history of exposure to the suspected risk factor:

**First**, select cases with the disease of interest and disease-free controls:

Cases	Controls
-------	----------



# Traditional Case-Control Study

**Then**, ascertain past history of exposure to the suspected risk factor:

**First**, select cases with the disease of interest and disease-free controls:


	Cases	Controls
Exposed		
Unexposed		

# Traditional Case-Control Study

**Then**, ascertain past history of exposure to the suspected risk factor:

**First**, select cases with the disease of interest and disease-free controls:

	Cases	Controls
Exposed	a	b
Unexposed	c	d
Total	a + c	b + d



**INCIDENCE RATES ARE NOT AVAILABLE IN A CASE-CONTROL STUDY**

---

<b>Design</b>	<b>Known variable at study's outset</b>	<b>Unknown variable the study wishes to ascertain</b>
Cohort	Presence of exposure to a suspected genetic or environmental risk factor	Incidence of the event (disease)

---

**Risk factor** ← → **Disease**

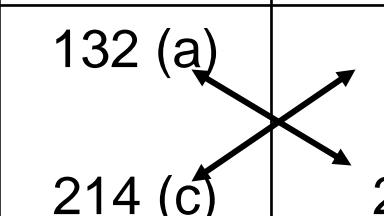
**For the traditional case-control study, the most important concept is that sampling of subjects for inclusion occurs at the end of a potential causal process**

<b>Design</b>	<b>Known variable at study's outset</b>	<b>Unknown variable the study wishes to ascertain</b>
Cohort	Presence of exposure to a suspected genetic or environmental risk factor	Incidence of the event (disease)
Case-control	Case-control status	Past exposure to suspected risk factor

## HOW TO MEASURE AN ASSOCIATION IN A CASE-CONTROL STUDY

Odds Ratios for the association maternal smoking and isolated clubfoot in the offspring, Atlanta, Georgia, 1968-80

Maternal smoking	Cases	Controls
Yes	132 (a)	866 (b)
No	214 (c)	2163 (d)
Total	346 (a+c)	3029 (b+d)



Relative Risk is the ratio of incidence rates/probabilities. Incidence cannot be calculated in case-control studies, for which the measure of association is the Odds Ratio:  $ad/bc$ .

Honein et al. Family history, maternal smoking, and clubfoot: an indication of gene-environment interaction. *Am J Epidemiol* 2000;152:658-65.

## HOW TO MEASURE AN ASSOCIATION IN A CASE-CONTROL STUDY

Odds Ratios for the association maternal smoking and isolated clubfoot in the offspring, Atlanta, Georgia, 1968-80

Maternal smoking	Cases	Controls	OR
Yes	132 (a)	866 (b)	$(132 \times 2163) \div (866 \times 2163) = 1.54$
No	214 (c)	2163 (d)	
Total	346 (a+c)	3029 (b+d)	

Relative Risk is the ratio of incidence rates/probabilities. Incidence cannot be calculated in case-control studies, for which the measure of association is the Odds Ratio:  $ad/bc$ .

Honein et al. Family history, maternal smoking, and clubfoot: an indication of gene-environment interaction. *Am J Epidemiol* 2000;152:658-65.

## HOW TO MEASURE AN ASSOCIATION IN A CASE-CONTROL STUDY

Odds Ratios for the association maternal smoking and isolated clubfoot in the offspring, Atlanta, Georgia, 1968-80

Maternal smoking	Cases	Controls	OR
Yes	132 (a)	866 (b)	$(132 \times 2163) \div 866 \times 2163 = 1.54$
No	214 (c)	2163 (d)	
Total	346 (a+c)	3029 (b+d)	

When the disease is relatively rare (e.g., <5%), the Odds Ratio is a good estimate of the Relative Risk

Honein et al. Family history, maternal smoking, and clubfoot: an indication of gene-environment interaction. *Am J Epidemiol* 2000;152:658-65.

# Observational Study Designs

- Cohort
- Case-control
  - Traditional (case-based)
  - Case-cohort:
    - A case-control study within a defined cohort



## **Example of case-cohort study**

Association between CMV antibodies and incident coronary heart disease (CHD) in the Atherosclerosis Risk in Communities (ARIC) Study

(Sorlie et al: *Arch Intern Med* 2000;160:2027-32)

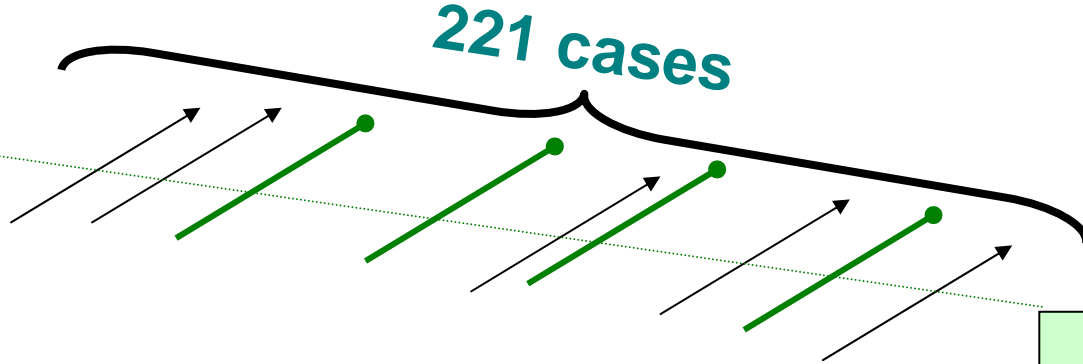
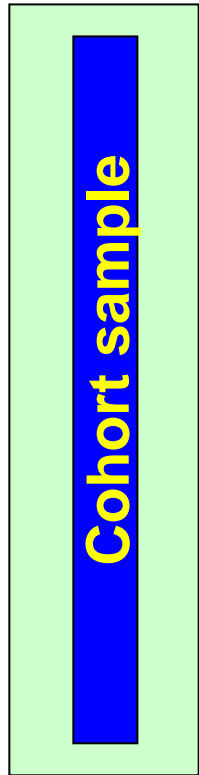
Cohort: 14,170 adult individuals (45-64 yrs at baseline) from 4 US communities (Jackson, Miss; Minneapolis, MN, Forsyth Co NC; Washington Co, MD), free of CHD at baseline.

Followed-up for up to 5 years.

# Case-cohort study

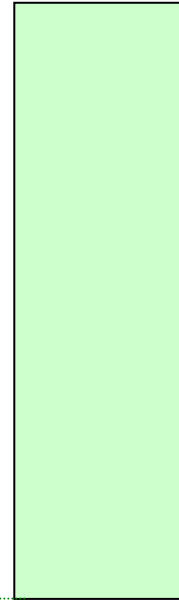
Random sample  
of 515 cohort subjects

N~14 170



Option 1= thaw serum samples  
of 14,000 persons, classify  
by CMV titer (+) or (-), and follow-  
up to calculate incidence in each  
group (exposed vs. unexposed)

**Option 2: Case-cohort study**



Initial  
pop

Final  
pop

Time (5 years)

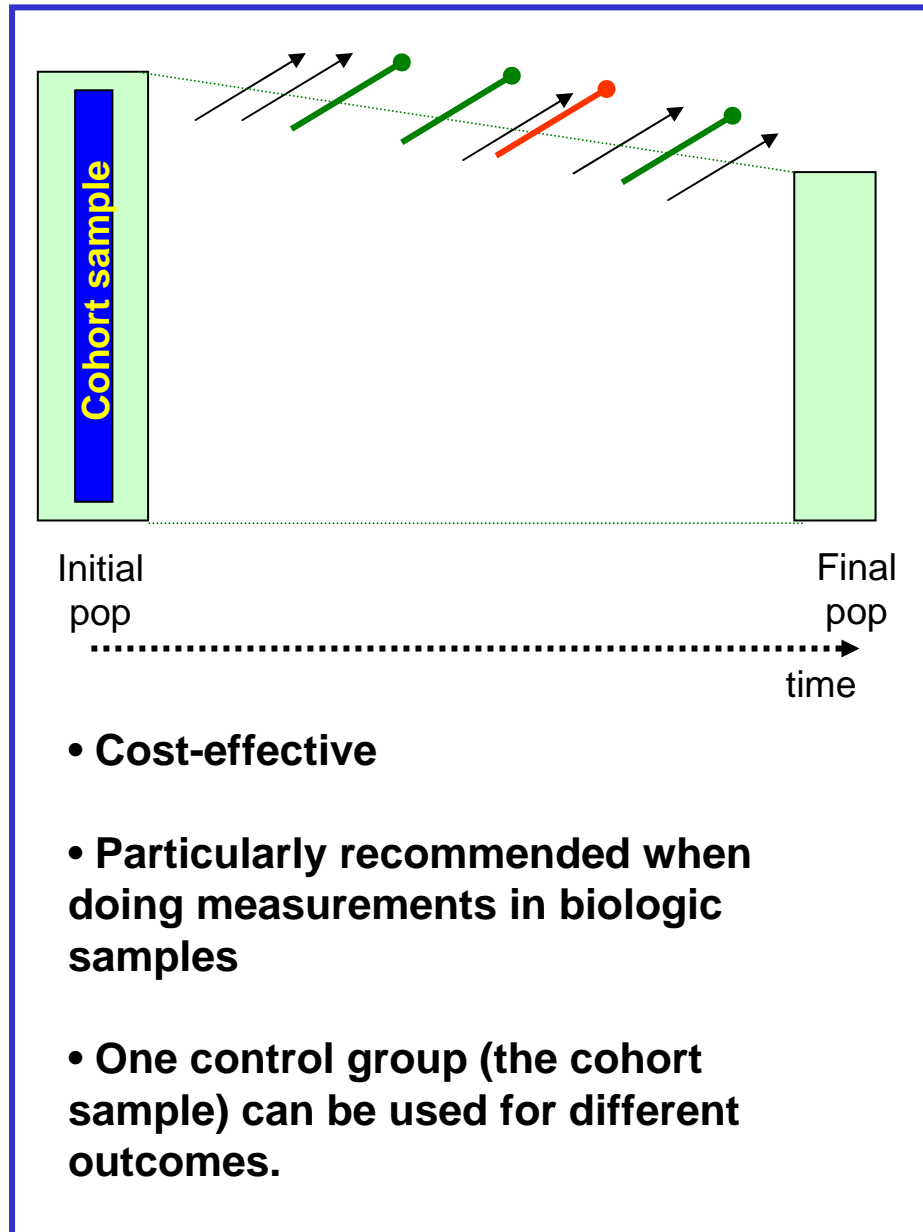
## Relative Risks of Coronary Heart Disease by Level of CMV Antibodies in the ARIC Study

CMV, P/N ratio	Relative Risk (95% CI)
0.0 – 1.9	1.00 (reference)
2.0 – 3.9	0.82 (0.40, 1.68)
4.0 – 5.9	0.90 (0.42, 1.90)
6.0+	1.89 (0.98, 3.67)

(Sorlie et al: *Arch Intern Med* 2000;160:2027-32)

Mathematically, the calculation of the odds ratio in a case-cohort study yields the relative risk

# Case-cohort Design



## “Effect Modification” or Interaction

Maternal smoking	Cases	Controls	OR
Yes	132 (a)	866 (b)	$(132 \times 2163) \div (866 \times 2163) = 1.54$
No	214 (c)	2163 (d)	

Family history of clubfoot	Maternal smoking	Cases	Controls	Stratified ORs
Yes	Yes	14	7	3.64
	No	11	20	
No	Yes	118	859	1.45
	No	203	2,143	

Honein et al. Family history, maternal smoking, and clubfoot: an indication of gene-environment interaction. *Am J Epidemiol* 2000;152:658-65.

## “Effect Modification” or Interaction

Maternal smoking	Cases	Controls	OR
Yes	132 (a)	866 (b)	$(132 \times 2163) \div (866 \times 2163) = 1.54$
No	214 (c)	2163 (d)	

Family history of clubfoot	Maternal smoking	Cases	Controls	Stratified ORs
Yes	Yes	14	7	<b>3.64</b>
	No	11	20	
No	Yes	118	859	<b>1.45</b>
	No	203	2,143	

Honein et al. Family history, maternal smoking, and clubfoot: an indication of gene-environment interaction. *Am J Epidemiol* 2000;152:658-65.

# Cohort Vs. Traditional Case-Control Studies

<b>ADVANTAGES AND DISADVANTAGES</b>	<b>COHORT</b>	<b>CASE-CONTROL</b>
Calculation of incidence rates and direct calculation of Relative Risks?	Yes	No. Odds Ratios estimate Rel. Risks for diseases with incidence <5%

# Cohort Vs. Traditional Case-Control Studies

<b>ADVANTAGES AND DISADVANTAGES</b>	<b>COHORT</b>	<b>CASE-CONTROL</b>
Calculation of incidence rates and direct calculation of Relative Risks?	Yes	No. Odds Ratios estimate Rel. Risks for diseases with incidence <5%
Length of study?	Long	Shorter



# Cohort Vs. Traditional Case-Control Studies

<b>ADVANTAGES AND DISADVANTAGES</b>	<b>COHORT</b>	<b>CASE-CONTROL</b>
Calculation of incidence rates and direct calculation of Relative Risks?	Yes	No. Odds Ratios estimate Rel. Risks for diseases with incidence <5%
Length of study?	Long	Shorter
Assessment of multiple exposures?	Yes	Yes

# Cohort Vs. Traditional Case-Control Studies

<b>ADVANTAGES AND DISADVANTAGES</b>	<b>COHORT</b>	<b>CASE-CONTROL</b>
Calculation of incidence rates and direct calculation of Relative Risks?	Yes	No. Odds Ratios estimate Rel. Risks for diseases with incidence <5%
Length of study?	Long	Shorter
Assessment of multiple exposures?	Yes	Yes
Assessment of multiple diseases (outcomes)?	Yes	Possible, but usually only one case group is studied

# Cohort Vs. Traditional Case-Control Studies

<b>ADVANTAGES AND DISADVANTAGES</b>	<b>COHORT</b>	<b>CASE-CONTROL</b>
Calculation of incidence rates and direct calculation of Relative Risks?	Yes	No. Odds Ratios estimate Rel. Risks for diseases with incidence <5%
Length of study?	Long	Shorter
Assessment of multiple exposures?	Yes	Yes
Assessment of multiple diseases (outcomes)?	Yes	Possible, but usually only one case group is studied
Ability to assess rare outcomes (e.g., Reye's syndrome, aplastic anemia)	Poor	Better

# Cohort Vs. Traditional Case-Control Studies

ADVANTAGES AND DISADVANTAGES	COHORT	CASE-CONTROL
Calculation of incidence rates and direct calculation of Relative Risks?	Yes	No. Odds Ratios estimate Rel. Risks for diseases with incidence <5%
Length of study?	Long	Shorter
Assessment of multiple exposures?	Yes	Yes
Assessment of multiple diseases (outcomes)?	Yes	Possible, but usually only one case group is studied
Ability to assess rare outcomes (e.g., Reye's syndrome, aplastic anemia)	Poor	Better
Ability to assess rare exposures (e.g., asbestos)	Greater	Poor

# Cohort Vs. Traditional Case-Control Studies

ADVANTAGES AND DISADVANTAGES	COHORT	CASE-CONTROL
Calculation of incidence rates and direct calculation of Relative Risks?	Yes	No. Odds Ratios estimate Rel. Risks for diseases with incidence <5%
Length of study?	Long	Shorter
Assessment of multiple exposures?	Yes	Yes
Assessment of multiple diseases (outcomes)?	Yes	Possible, but usually only one case group is studied
Ability to assess rare outcomes (e.g., Reye's syndrome, aplastic anemia)	Poor	Better
Ability to assess rare exposures (e.g., asbestos)	Greater	Poor
Cost?	+++	+

# Cohort Vs. Traditional Case-Control Studies

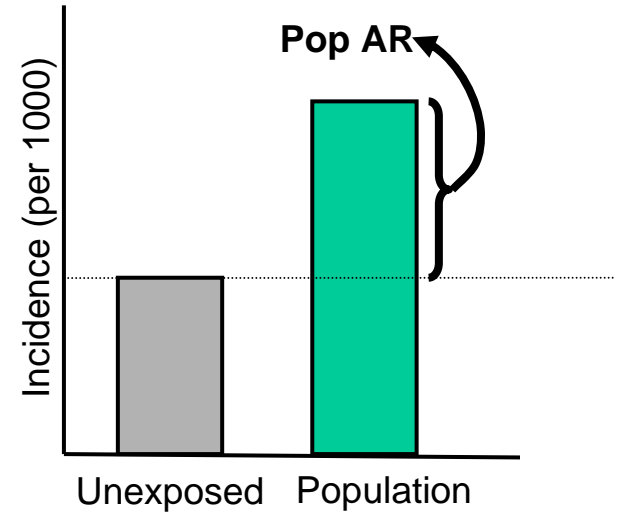
ADVANTAGES AND DISADVANTAGES	COHORT	CASE-CONTROL
Calculation of incidence rates and direct calculation of Relative Risks?	Yes	No. Odds Ratios estimate Rel. Risks for diseases with incidence <5%
Length of study?	Long	Shorter
Assessment of multiple exposures?	Yes	Yes
Assessment of multiple diseases (outcomes)?	Yes	Possible, but usually only one case group is studied
Ability to assess rare outcomes (e.g., Reye's syndrome, aplastic anemia)	Poor	Better
Ability to assess rare exposures (e.g., asbestos)	Greater	Poor
Cost?	+++	+
Probab. of selection/information bias?	+	+++

# Cohort Vs. Traditional Case-Control Studies

ADVANTAGES AND DISADVANTAGES	COHORT	CASE-CONTROL
Calculation of incidence rates and direct calculation of Relative Risks?	Yes	No. Odds Ratios estimate Rel. Risks for diseases with incidence <5%
Length of study?	Long	Shorter
Assessment of multiple exposures?	Yes	Yes
Assessment of multiple diseases (outcomes)?	Yes	Possible, but usually only one case group is studied
Ability to assess rare outcomes (e.g., Reye's syndrome, aplastic anemia)	Poor	Better
Ability to assess rare exposures (e.g., asbestos)	Greater	Poor
Cost?	+++	+
Probab. of selection/information bias?	+	+++
Time sequence (exposure→outcome)	Clear	Can be unclear

- **Population attributable risk:**

The excess risk in the population that can be attributed to a given risk factor.



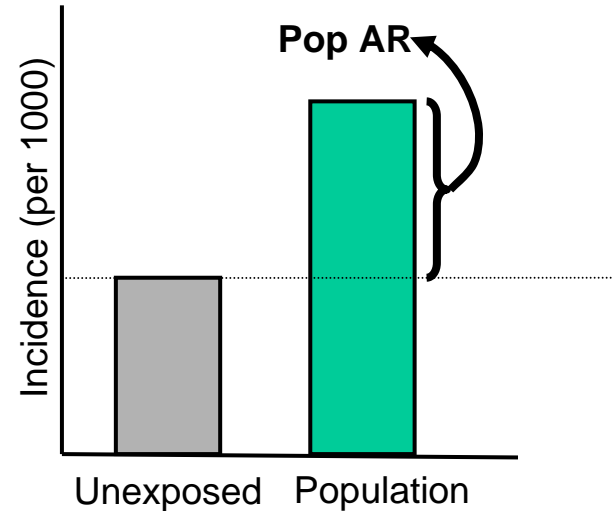


Unadjusted Relative Risk for Diabetes in Relation to Coronary Heart Disease, ARIC Cohort Study, Women

	Sample Size	No. Events	Rate/100	RR
Diabetes				
Yes	614	35	5.7	<b>6.3</b>
No	6 675	61	0.9	<b>1.0</b>

(Chambless et al, *Am J Epidemiol* 1997;146:483-94)

• **Population attributable risk:**  
The excess risk in the population that can be attributed to a given risk factor.



Levin's formula:

$$\% \text{Pop AR} = \frac{p_e(RR - 1)}{p_e(RR - 1) + 1} \times 100$$

(Levin: *Acta Un Intern Cancer* 1953;9:531-41)

Prevalence of diabetes =  $614/7289 = 0.084$

$$\% \text{PopAR} = \frac{0.084(6.3 - 1)}{0.084(6.3 - 1) + 1} \times 100 = 30.8\%$$

Levin's formula can be only used for unadjusted data. More complex formulas are available for adjusted relative risks