Biost 517 / Biost 514 Applied Biostatistics I / Biostatistics I

Timothy A. Thornton, Ph.D.
Associate Professor of Biostatistics
University of Washington

Lecture 12:

Cross Sectional Studies, Cohort Studies, and Case-Control Studies; Validity of Risk Differences, Risk Ratios, and Odds Ratios by Study Design

II. Common Study Designs

- Interventional trials
- Cohort studies
- Case-control studies
- Cross-sectional studies

Interventional Trials

- Subjects assigned to some intervention
  - Ideally controlled, randomized, blinded
- Followed longitudinally for some outcome
- Efficient for examining
  - Common outcomes
- The gold standard for establishing cause and effect

Interventional Trial: important terminology

- “randomized controlled double-blind clinical trial”
  - “randomized” Patients randomly assigned into treatment groups
  - “controlled” There is some control group
    - Placebo
    - Standard of care
  - “double blind”
    - patients do not know what treatments they are receiving
    - the clinicians assessing the outcome do not know either

Randomized trials: strengths

1. Duration and intensity of treatment is under the control of the investigators
2. Exposure is known to precede outcome
3. Best protection against **confounding. A confounder** is an extraneous variable that confuses the association between a predictor of interest and an outcome variable if not properly accounted for. We will cover confounding in the near futre
   - Randomization implies that, on average, treatment assignment is not associated with other factors

Randomized trials: issues

1. Many exposures can’t be studied by random assignment
   - especially harmful exposures
2. Difficult to study exposures with a long latent period
3. Difficult to study rare diseases or outcomes
4. “blinding” may not be practical
   - e.g., surgery vs. physical therapy

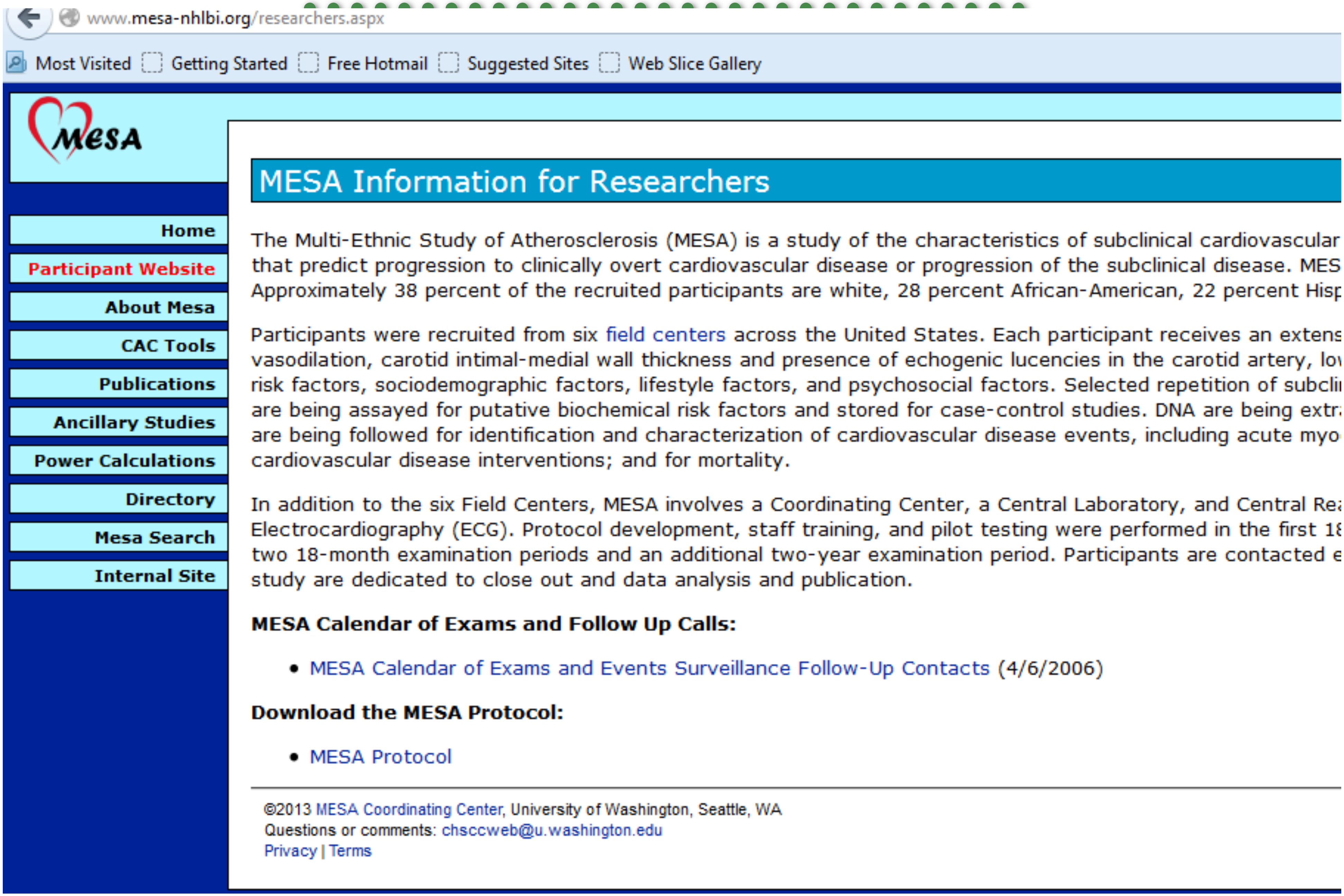
Randomized trials: issues

5. Loss to follow-up
   - potential solution: restrict study population
6. Non-compliance with assigned treatment
7. Potential lack of generalizability
   - artificial nature of the study, restricted study population
   - different "type" of patients volunteer for trials

Cohort Studies

- Group or groups followed longitudinally for outcome(s)
  - Prospectively into the future, or
  - Retrospectively since some defining event
    - e.g., since being born in a particular hospital in a particular year
- Efficient for examining
  - Common outcomes
  - Many different outcomes for same exposure
  - Associations (not cause and effect)
  - Estimate incidence within risk factor groups
    - Cannot estimate prevalence of a risk factor that was part of the sampling scheme

MESA cohort study: www.mesa-nhlbi.org



MESA Information for Researchers

The Multi-Ethnic Study of Atherosclerosis (MESA) is a study of the characteristics of subclinical cardiovascular that predict progression to clinically overt cardiovascular disease or progression of the subclinical disease. MES Approximately 38 percent of the recruited participants are white, 28 percent African-American, 22 percent Hisp

Participants were recruited from six field centers across the United States. Each participant receives an extens vasodilation, carotid intimal-medial wall thickness and presence of echogenic lucencies in the carotid artery, lo risk factors, sociodemographic factors, lifestyle factors, and psychosocial factors. Selected repetition of subcli are being assayed for putative biochemical risk factors and stored for case-control studies. DNA are being extr are being followed for identification and characterization of cardiovascular disease events, including acute myo cardiovascular disease interventions; and for mortality.

In addition to the six Field Centers, MESA involves a Coordinating Center, a Central Laboratory, and Central Re Electrocardiography (ECG). Protocol development, staff training, and pilot testing were performed in the first 1 two 18-month examination periods and an additional two-year examination period. Participants are contacted e study are dedicated to close out and data analysis and publication.

MESA Calendar of Exams and Follow Up Calls:

- MESA Calendar of Exams and Events Surveillance Follow-Up Contacts (4/6/2006)

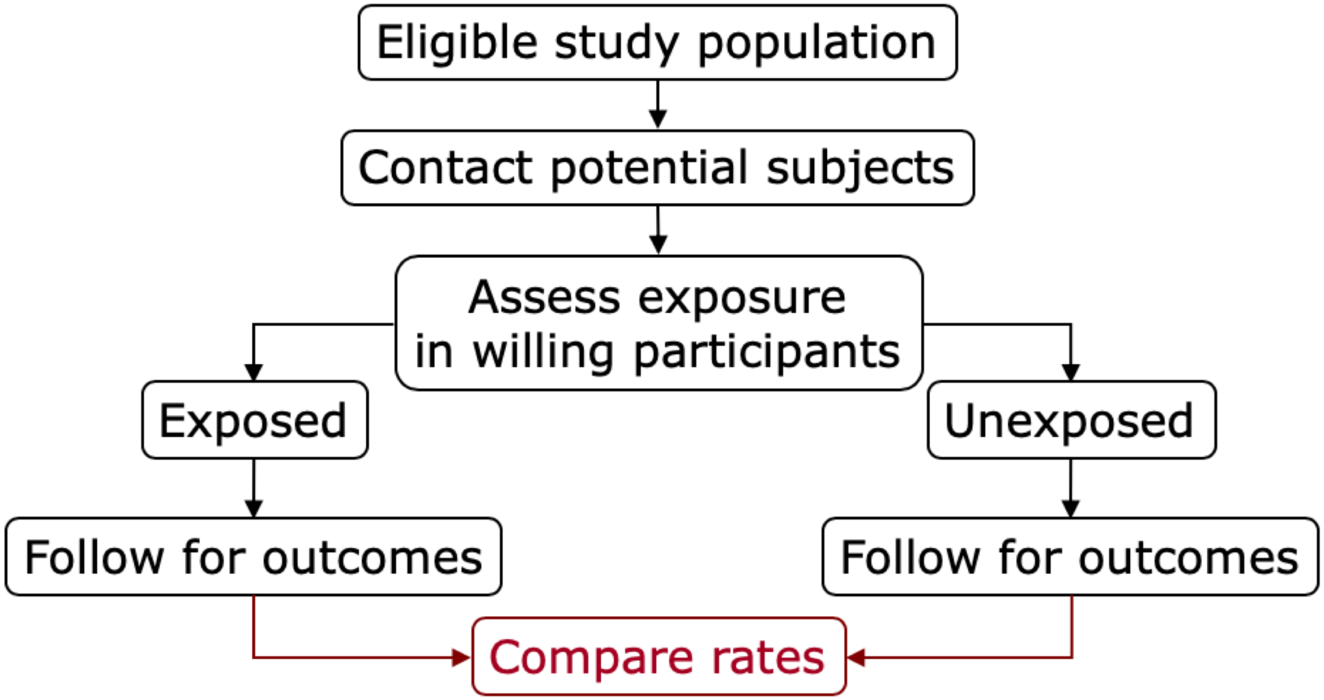
Download the MESA Protocol:

- MESA Protocol

©2013 MESA Coordinating Center, University of Washington, Seattle, WA
Questions or comments: chsccweb@u.washington.edu
Privacy | Terms

Cohort Studies: How they typically are conducted

- Conceptually similar to interventional trial, but “exposure” is NOT assigned by the investigator
- “Observational” version of an interventional trial


Eligible study population

Contact potential subjects

Assess exposure in willing participants

Exposed

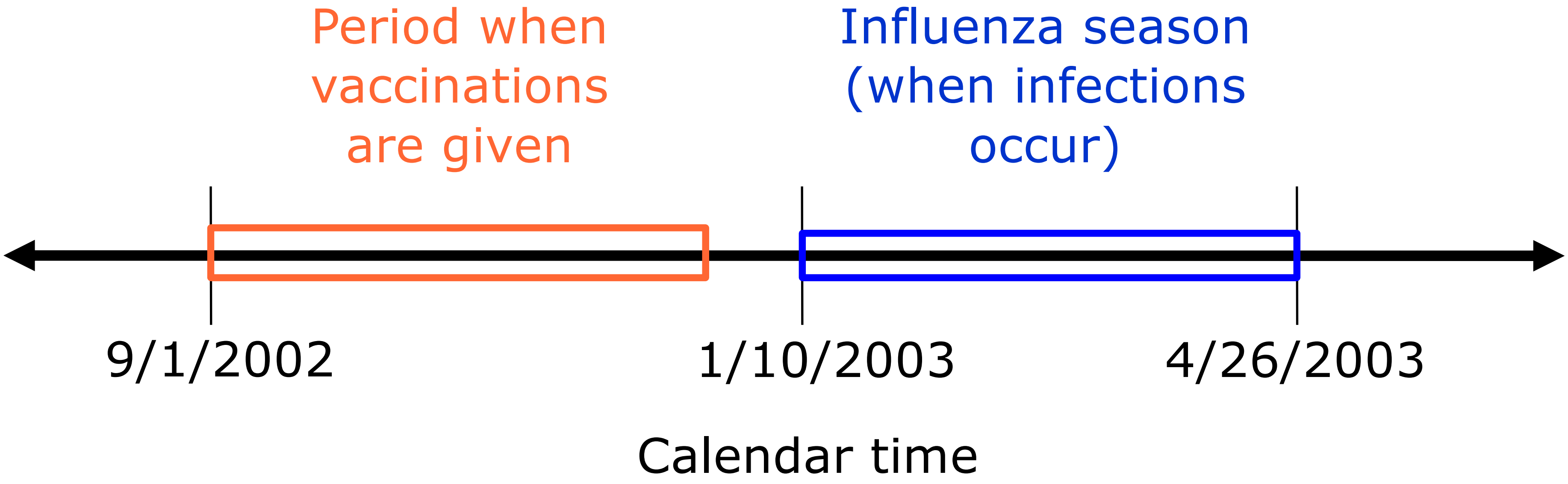
Unexposed

Follow for outcomes

Follow for outcomes

Compare rates

Prospective vs retrospective



Period when vaccinations are given

Influenza season (when infections occur)

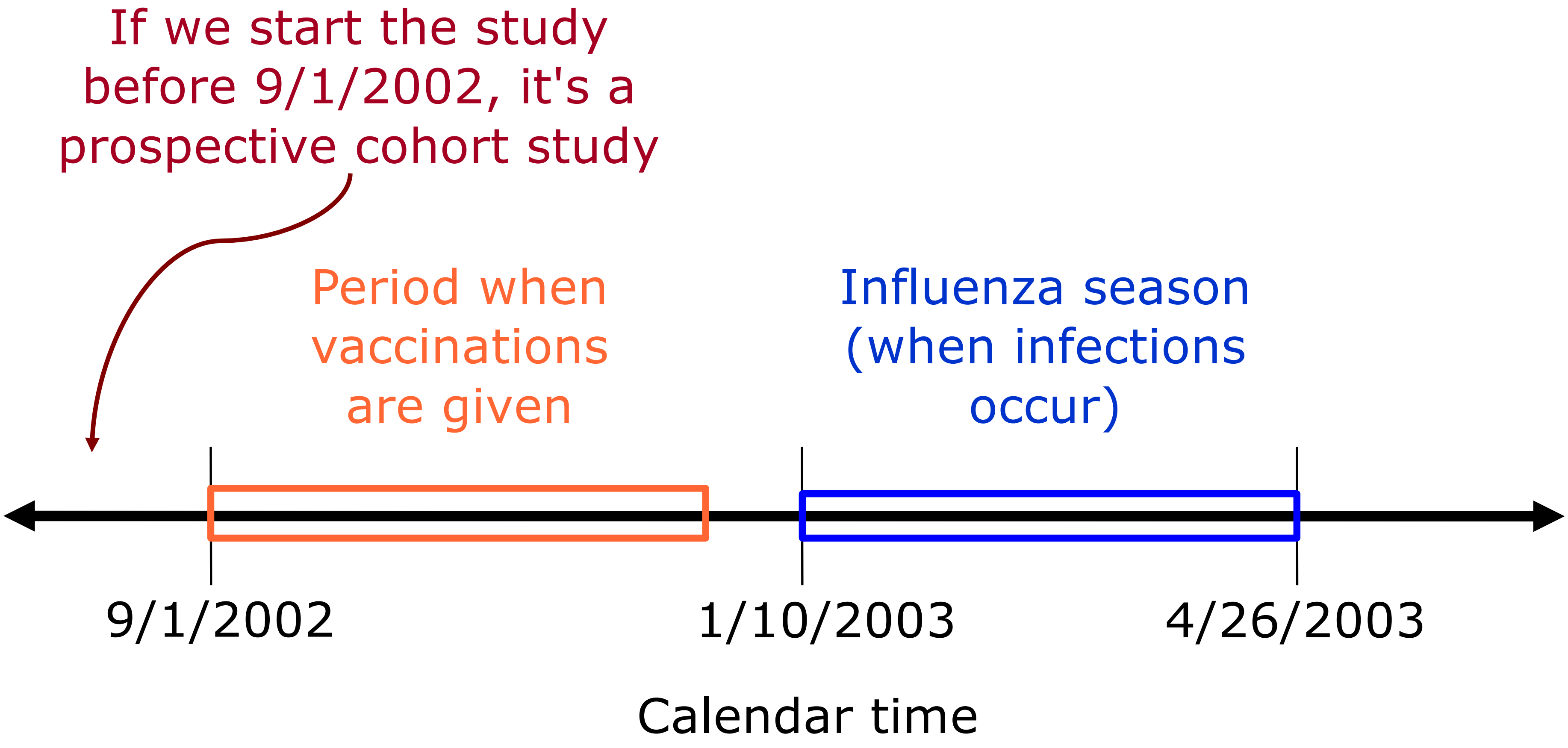
9/1/2002

1/10/2003

4/26/2003

Calendar time

Prospective vs retrospective

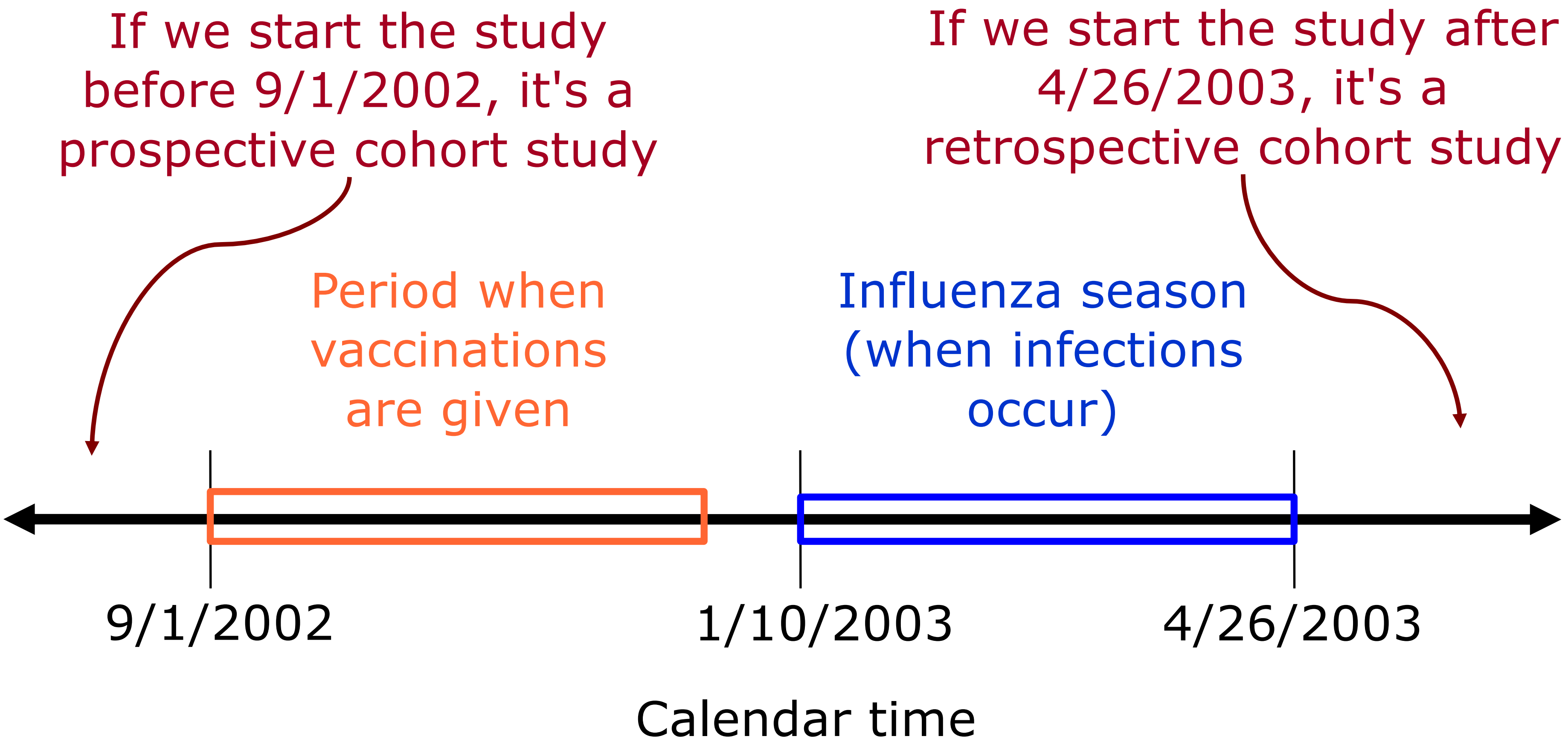


If we start the study before 9/1/2002, it's a prospective cohort study

Period when vaccinations are given

Influenza season (when infections occur)

Prospective vs retrospective



If we start the study before 9/1/2002, it's a prospective cohort study

If we start the study after 4/26/2003, it's a retrospective cohort study

Period when vaccinations are given

Influenza season (when infections occur)

9/1/2002

1/10/2003

4/26/2003

Calendar time

Cohort studies: strengths

1. Exposure is known to precede disease
2. Can study disease risk factors and exposures that aren’t modifiable by investigators
3. Can investigate multiple outcomes
4. Potentially better generalizability than randomized trials

Cohort studies: potential problems

1. Differences between exposed and unexposed subjects (confounding)

   Solution: Measure and adjust for differences

2. Biased outcome assessment

   Solution: Blinded outcome assessment

3. Loss to follow-up in prospective study

   Solution: Restricted study population

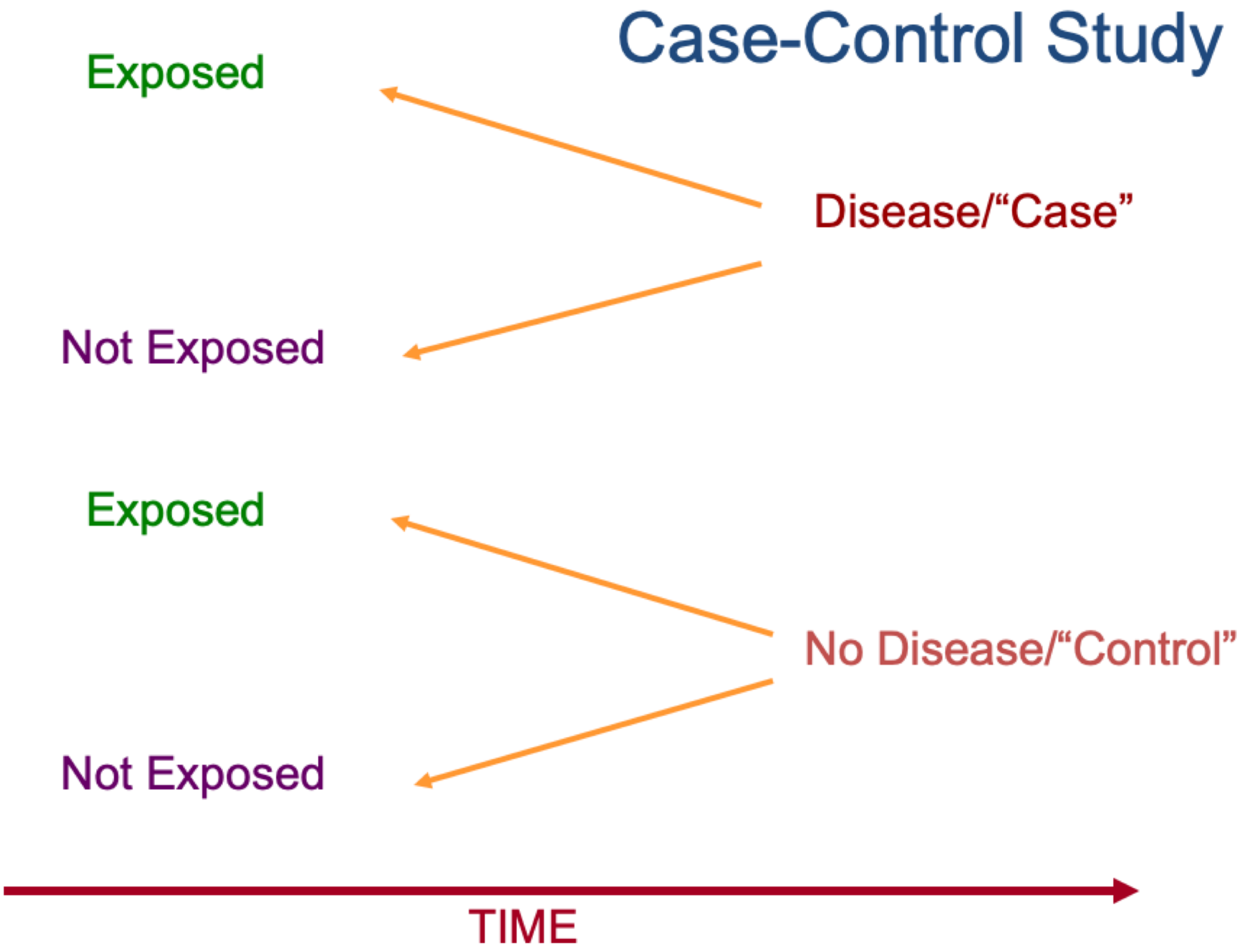
Cohort studies: general limitations

1. Difficult to study rare diseases
2. Difficult to study exposures with a long latent period
3. Prospective studies can be expensive and time consuming
4. Retrospective studies usually must rely on data that have been collected for non-scientific purposes

Case-control Studies

- Sample based on the outcome/disease
- Retrospective
- *Nature Genetics*, October 2014
  - "A new study compares the copy number variants (CNVs) in 29,085 children with developmental delay to those in 19,584 healthy controls…"

Case-Control Study:



Case-Control Study

Exposed

Disease/“Case”

Not Exposed

Exposed

No Disease/“Control”

Not Exposed

TIME

Case-Control Studies

- Identify individuals with disease/outcome of interest and a comparable control group at risk for disease/outcome

- Look back in time to determine proportion of cases and controls that were exposed and non-exposed

- Examples of cases: cancer cases, adverse event, diabetic patients, emphysema patients

- Examples of exposure: medication use, environmental exposure, dietary choices

Identifying Cases

- Ideally, random sample of everyone who develops disease or all diseased in sample population
  - Example: all breast cancer cases diagnosed in 13 WA counties during 1990-2008
- Type of cases
  - Incident (newly diagnosed) – generally preferred (next slide)
  - Prevalent (all existing cases – new and old)

Incident versus Prevalent Cases

- Including prevalent cases in the “case sample” means that patients with a long course of disease tend to be over-represented.
  - By definition, those with a short duration of disease stop being “cases” because of either recovery or death.
- “Unless we can justify the assumption that the exposure being studied is not associated with recovery or survival, every effort should be made to limit recruitment to incident cases.”
  - Cancer Epidemiology: Principles and Methods, I.D.S. Silva, Ed.

Selection of Controls

- Controls should be selected from same population that gives rise to cases
  - This can be difficult to characterize
- Controls should be similar to cases in all respects other than having disease
  - Patient characteristics, co-morbidities, etc.

Case-Control Studies

- Characterize prior exposures
  - Longitudinal study into the past
  - How to handle time in exposure?
    - ever / never exposed, cumulative exposure, time since exposure
- Efficient for examining
  - Rare outcomes
  - Many different risk factors for same outcome
  - Estimate prevalence of exposure by disease
    - Cannot estimate prevalence of disease

Cross-sectional Studies

- Surveys of subjects sampled from a population
- Real or event time
  - "Real time" = "calendar time"
  - "Event time" = when some event happens
    - birth, marriage, diagnosis, treatment, death
- Efficient for examining
  - Common outcomes and risk factors
  - Associations (not cause and effect)
  - Can estimate prevalence of risk factors and outcomes
    - Overall and within groups

Cross-Sectional Studies

- “Snapshot” at a particular point in time
- Quick, convenient, often inexpensive compared to other types of studies
- Can look at multiple exposures and outcomes

Cross-Sectional Studies: Limitations

- Identifies prevalent, not incident, cases
- May under-represent disease cases with short duration

Choice of Summary Measure / Contrast

- Interplay of
  - How we want to eventually use results of the analysis
    - Which variable do we ideally want to condition on?
  - And the way we sampled our data
    - Were sample sizes in any subpopulations fixed by design?
- Example: Problem of scientific interest is to know if radon exposure increases risk getting lung cancer.
- Cohort studies often cost too much and may not be feasible in terms of the time needed to collect the data:
  - Identify a group of people who are exposed to radon and group of people who are not exposed
  - Follow them until enough of them get lung cancer so that the number of events is large enough to statistically evaluate risk

Choice of Summary Measure / Contrast

- Even though I want to be able to look at the group of people who have radon (the exposure) and estimate the probability of getting lung cancer, I may have to sample my data with a **case-control study design**
- In the case-control study design:
  - Fix the number of individuals who got lung cancer and the number of individuals who did not get cancer in my data
  - I then look back to see how many were exposed to radon.
- I desire to scientifically to condition on radon exposure and talk about the probability of getting cancer. My case-control sampling scheme, however, dictates that all I am allowed to do is condition on disease status and talk about the distribution of radon exposure.
- So there is this clash between these two. Now in general, this would dictate how I would have to do Risk Difference or Risk Ratio regression. But as I am going to demonstrate, when I am performing odds ratio regression, it is ok to answer the question either way. And because it is ok to answer the question either way, this makes odds ratios more desirable when doing case-control sampling.

Case Control Studies: Equivalence of Odds Ratios for Exposure and Disease

- Let ***a*** and ***b*** be the counts for the Cases and Controls, respectively, that were exposed. Let ***c*** and ***d*** be the counts for the number cases and controls, respectively, that were not exposed.

	Cases	Controls
Exposed	a	b
Unexposed	c	d

Odds and Odds Ratio of Disease

	Cases	Controls
Exposed	a	b
Unexposed	c	d

Odds of being a case for exposed group$=\dfrac{a}{b}$

Odds of being a case for unexposed group$=\dfrac{c}{d}$

Odds Ratio of being a case for exposed and unexposed group$=\dfrac{a/b}{c/d}=\dfrac{ad}{cb}$

Odds and Odds Ratio of Exposure

	Cases	Controls
Exposed	a	b
Unexposed	c	d

$$\text{Odds of exposure for cases } = \frac{a}{c}$$

$$\text{Odds of exposure for controls } = \frac{b}{d}$$

$$\text{OR of exposure for cases and controls } = \frac{a/c}{b/d} = \frac{ad}{cb}$$

Choice of Summary Measure / Contrast

- Want to condition on exposures, but use case-control sampling with rare disease ➔ perhaps prefer using OR
  - In general if we constrain sample sizes for diseased vs non-disease, I really should only talk about the distribution of exposures condition on diseases: *Pr(Exposure | Disease)*
  - As we just demonstrated, the odds ratio is the exception for a summary measure, because the odds ratio of exposure conditioned on disease status is equal to the odds ratio of disease when conditioning on exposure status.
  - So we can talk about the odds ratio in either direction!

Cohort Studies

- Scientific interest:
  - Distribution of “effect” across groups defined by “cause”
  - E.g., how does risk of lung cancer differ by smoking behavior

- Common sampling schemes
  - Cohort study: Sample by exposure
    - Sample 1000 smokers, 1000 nonsmokers
    - Estimate risk of lung cancer in exposure groups

Case-Control Studies

- Case-control study: Sample by outcomes
  - Sample 1000 lung cancer patients, 1000 controls
  - In general I should be estimating prevalence of smoking in diagnosis groups, e.g., proportion (or odds) of smokers among people with or without cancer
    - In a case-control study, I cannot estimate the proportion of cancer outcomes or odds of cancer defined by their smoking status: I didn’t sample that way.
    - I can only estimate the proportion or odds of smoking by diagnostic group.
    - However, I can look at the odds ratio, even when I cannot estimate the individual odds, as we will show.

Use of Odds Ratios

- Cohort study
  - Odds of cancer among smokers : odds of cancer among nonsmokers
- Case-control study
  - Odds of smoking among cancer : odds of smoking among individuals without cancer
- Mathematically, the two odds ratios are the same
  - Hence, when using case-control sampling, it is valid to estimate either odds ratio

Example: Two Sample Studies

- Investigate association between mortality and smoking in a population of elderly adults
  - Death within 4 years of some “sentinel event”
  - Smoking behavior current at time of the “sentinel event”
- Sampling schemes that might be considered
  - View this as a cross-sectional sampling of 4,994 subjects
    - Can estimate either distribution: smoking conditional on death status, or estimate mortality within groups defined by smoking status
  - Cohort study: Randomly sampled 400 smokers and 1,200 nonsmokers
  - Case-control study: Randomly sampled of 300 individuals who died within 4 years of sentinel event and 900 controls alive 4 years after the sentinel event
  - We chose sample size for individuals who smoked and did not smoke for the cohort. Similarly, we chose the size of those who died and did not die. For the cross-sectional study, nature chose the sizes of smokers/no-smokers and death/no-death within 4 years.

Cross-sectional Study (N_{Tot} = 4,994)

- Valid estimates: Mortality conditioning on smoking behavior
- Valid estimates: *Smoking behavior conditioning on mortality*

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="2">Death w/in 4 Yr</td><td colspan="2" rowspan="2">Pr (Dth = 1 | Smk)</td><td rowspan="2">Odds (Dth = 1 | Smk)</td></tr>
<tr><td>0</td><td>1</td></tr>
<tr><td rowspan="2">Smoking</td><td>0</td><td>3,966</td><td>424</td><td colspan="2">0.0966</td><td>0.1069</td></tr>
<tr><td>1</td><td>533</td><td>71</td><td colspan="2">0.1175</td><td>0.1332</td></tr>
<tr><td colspan="2">Pr (Smk = 1 | Dth)</td><td>0.1185</td><td>0.1434</td><td>RD
Dth | Smk: .0210
Smk | Dth: .0250</td><td>RR
Dth | Smk: 1.217
Smk | Dth: 1.211</td><td></td></tr>
<tr><td colspan="2">Odds (Smk = 1 | Dth)</td><td>0.1344</td><td>0.1675</td><td></td><td></td><td>OR
Dth | Smk: 1.246
Smk | Dth: 1.246</td></tr>
</table>

Cohort Study (N_{Smk}= 400; N_{NS}= 1,200)

- Valid estimates: Mortality conditioning on smoking behavior
- Valid estimates: *Smoking behavior conditioning on mortality*

		Death w/in 4 Yr		Pr (Dth = 1 | Smk)		Odds (Dth = 1 | Smk)
		0	1			
Smoking	0	**1,090**	**110**	**0.0917**		**0.1009**
	1	**358**	**42**	**0.1050**		**0.1173**
Pr (Smk = 1 | Dth)		~~0.2472~~	~~0.2763~~	**RD** **Dth | Smk: .0133** ~~Smk | Dth: .0291~~	**RR** **Dth | Smk: 1.145** ~~Smk | Dth: 1.118~~	
Odds (Smk = 1 | Dth)		~~0.3284~~	~~0.3818~~			**OR** **Dth | Smk: 1.163** ***Smk | Dth: 1.163***

Case-Control Study (N_{Die}= 300; N_{Surv}= 900)

- Valid estimates: Mortality conditioning on smoking behavior
- Valid estimates: *Smoking behavior conditioning on mortality*

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="2">Death w/in 4 Yr</td><td colspan="2" rowspan="2">Pr (Dth = 1 | Smk)</td><td rowspan="2">Odds (Dth = 1 | Smk)</td></tr>
<tr><td>0</td><td>1</td></tr>
<tr><td rowspan="2">Smoking</td><td>0</td><td>783</td><td>259</td><td colspan="2">0.2486</td><td>0.3308</td></tr>
<tr><td>1</td><td>117</td><td>41</td><td colspan="2">0.2595</td><td>0.3504</td></tr>
<tr><td colspan="2">Pr (Smk = 1 | Dth)</td><td>0.1300</td><td>0.1367</td><td>RD
Dth | Smk: .0109
Smk | Dth: .0067</td><td>RR
Dth | Smk: 1.044
Smk | Dth: 1.051</td><td></td></tr>
<tr><td colspan="2">Odds (Smk = 1 | Dth)</td><td>0.1494</td><td>0.1583</td><td></td><td></td><td>OR
Dth | Smk: 1.059
Smk | Dth: 1.059</td></tr>
</table>

Case-Control Study (N_{Die}= 300; N_{Surv}= 900)

- Valid estimates: Mortality conditioning on smoking behavior
- Valid estimates: *Smoking behavior conditioning on mortality*

<table>
<tr><th colspan="2" rowspan="2"></th><th colspan="2">Death w/in 4 Yr</th><th colspan="2" rowspan="2">Pr (Dth = 1 | Smk)</th><th rowspan="2">Odds (Dth = 1 | Smk)</th></tr>
<tr><th>0</th><th>1</th></tr>
<tr><td rowspan="2"><b>Smoking</b></td><td>0</td><td><b>783</b></td><td><b>259</b></td><td colspan="2">~~0.2486~~</td><td>~~0.3308~~</td></tr>
<tr><td>1</td><td><b>117</b></td><td><b>41</b></td><td colspan="2">~~0.2595~~</td><td>~~0.3504~~</td></tr>
<tr><td colspan="2"><b>Pr (Smk = 1 | Dth)</b></td><td><i>0.1300</i></td><td><i>0.1367</i></td><td><b>RD</b>
~~Dth | Smk: .0109~~
<b><i>Smk | Dth: .0067</i></b></td><td><b>RR</b>
~~Dth | Smk: 1.044~~
<b><i>Smk | Dth: 1.051</i></b></td><td></td></tr>
<tr><td colspan="2"><b>Odds (Smk = 1 | Dth)</b></td><td><i>0.1494</i></td><td><i>0.1583</i></td><td></td><td></td><td><b>OR</b>
<b>Dth | Smk: 1.059</b>
<b><i>Smk | Dth: 1.059</i></b></td></tr>
</table>

Take Home Message 1

- The corresponding valid estimates from each study design are estimating the same quantity
  - E.g., both the cross-sectional and cohort studies can be used to estimate the population *Pr[Die w/in 4 years* | *Smoke]*
- I created a single cohort design and a single case-control design by sampling from the cross-sectional design.
  - There is of course less precision in those derived designs, because the sample sizes are smaller
  - Furthermore, the cross-sectional design did not have all that much precision
    - The inference from the cross-sectional study estimated an odds ratio of 1.246, with a 95% CI of 0.95 to 1.63
    - The estimated OR from the cohort and case-control studies (which were 1.163 and 1.059, respectively) were consistent with that lack of precision

Take Home Message 2

- All study designs are estimating the odds ratio comparing the odds of death within 4 years for smokers to the odds of death within 4 years for nonsmokers
  - The cross-sectional and cohort studies can do this directly
  - The case-control study can do this indirectly from a scientific standpoint
    - But because this is true scientifically, and because the OR is mathematically the same in either direction, we can actually fit the “reverse” logistic regression model and get the same answer for the slope (though the intercept in that model is not estimating a population-based odds)
- This property is an advantage of looking at OR, because with rare events, case-control sampling is more feasible and economical

Take Home Message 3

- The previous discussions related to choice of analysis method by study design as it relates to the quantification of the association
- However, if we are only interested in establishing an association, then all we need to consider is asses p values:
  - Comparing distribution of Y across subgroups defined by X versus comparing distribution of X across subgroups defined by Y
  - P values tend to be quite similar no matter which analysis we use
  - For odds ratios, the results are identical due to the invariance property of odds ratios

OR Interpretation in Case-Control Studies

- The odds ratio is easily interpreted when trying to investigate rare events
  - Odds = prob / (1 – prob)

  - Rare event: (1 – prob) is approximately 1
    - Odds is approximately the probability
    - Odds ratio is approximately the risk ratio
      - Risk ratios are easily understood

- Case-control studies are typically used when events are rare

- Note that in the previous example, the probability of death was on the order of 10% in the cross-sectional study, so the OR and the RR are only approximately equal.

Final thought: Be Careful with Comparing Ratios

- How close are two ratios?
  - 0.20 and 0.25 VERSUS 5.0 and 4.0 ?
  - 0.10 and 0.15 VERSUS 10.0 and 6.7 ?

- For inference on an association from the ratios, they are identical
- We might tend to consider a bigger difference when two ratios are each > 1 than when they are each < 1
  - "But that would be wrong."