Design of Epidemiologic Studies

Laura Lee Johnson, Ph.D.
Statistician
National Center for Complementary and Alternative Medicine

Introduction to the Principles and Practice of Clinical Research

Monday, October 22, 2007

Epidemiology & Biostatistics

* Design of Epidemiologic Studies
* Study Development
* Randomization
* Hypothesis Testing
* Sample Size and Power
* Survival Analysis

Objectives

* Intuitive understanding of some study designs and statistics used in clinical research
* Understand and perform some simple but useful analyses & sample size calculations
* Little language to work with statisticians and epidemiologists

Housekeeping 1

* Six lectures
  - Overlaps: complement and help lead you a bit deeper with each discussion
* Exam questions based on lectures
  - Class book will help but does not have all the answers
* Appendices, book and software lists are for your reference

Housekeeping 2

* See book errata sheet for biostat chapters
* Slides: color (not needed) vs. bwprint
* 90 min lectures

JAMA Need To Know

* Sept 5, 2007 **Windish, D. M. et al. JAMA 2007;298:1010-1022**
* **16 questions**
* **Answer them before/after lectures**
* **All levels in this class**
  - **PhD in epi or biostat: please help your peers**

Analysis Follows Design

Questions - Hypotheses - Experimental Design – Samples - Data - Analyses -Conclusions

* Take all of your design information to a statistician early and often
  - Guidance
  - Assumptions

Do Not Confuse

* Association
* Causality
* Confounding

Objectives:
Design of Epi Studies Lecture

* What is Epidemiology?
* Along with the next few lectures, what are several types of studies

Outline

* *What is Epidemiology?*
- Vocabulary (1)
- Types of studies
- Questions

What is Epidemiology?

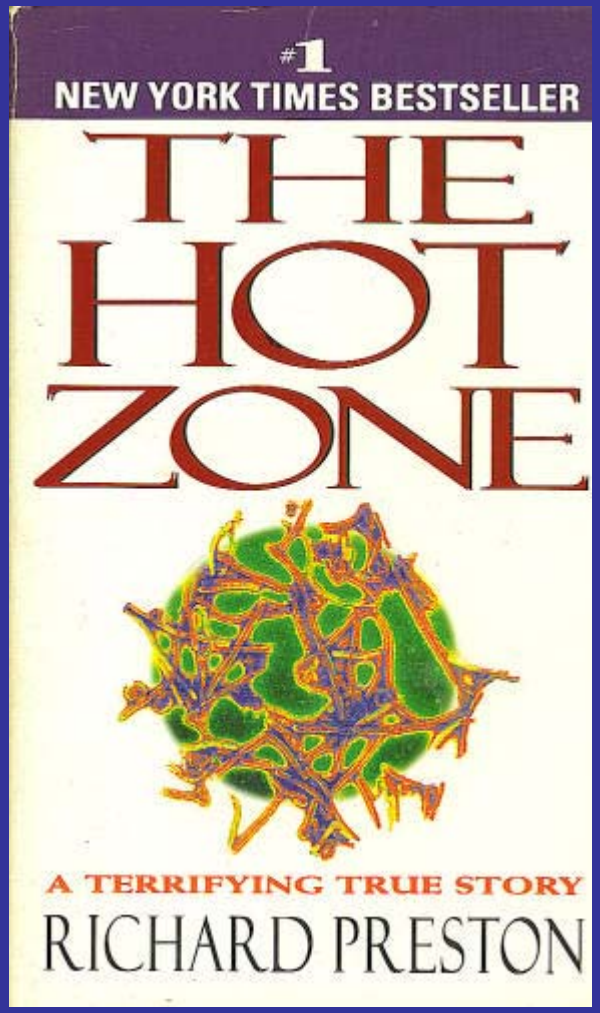
* Study of the distribution and determinants of disease and injury in human populations
  - Human disease does not occur at random
  - Human disease has causal and preventive factors that can be identified through systematic investigation of different populations or subgroups of individuals within a population

    * Hennekens and Buring, 1987

What is Epidemiology?

* Studying epidemics
  - EIS (Epidemic Intelligence Service) at CDC

EIS? Think Outbreak
Epi - Epidemic



What is Epidemiology?

* Studying epidemics
  - EIS (Epidemic Intelligence Service) at CDC
* Big cohort studies
  - Nurses Health Study
* Many things in between
* Considered (by some) cornerstone of Public Health Research

Big Studies?
Are They All Refuted?

* All the epidemiologic studies prior to Women's Health Initiative (WHI)
  - They did not agree, really
  - Now with new stat methods, well, hindsight was 20/20
* Prevention?
* NYT Magazine Sept 16, 2007

What is Hard

* Measure many things
* Measure each thing many different ways
* Measure each of those VERY accurately
  - Often
  - Do not lose any data
  - Same way every time
* What you don’t know and don’t measure

Epidemiology and Hypotheses

* Epidemiology is hypothesis generating evidence
  - Like circumstantial evidence in court?
* May be the only information outside of the laboratory
* Fundamental limitation
  - Distinguish associations
  - CANNOT inherently determine causation

Generating Hypotheses

* Epidemiology
* Clinician experience/observation
* Out of thin air

Causal Inference in Observational Studies: Epidemiologic Criteria

A. Statistical significance
B. Strength of association (odds ratio, relative risk)
C. Dose-response relationships
D. Temporal sequence
E. Consistency of the association (internal "validity")
F. Replication of results (external validity)
G. Biological plausibility
H. Experimental evidence

Outline

- What is Epidemiology?
* *Vocabulary (1)*
- Types of studies
- Questions

Words I Might Use

* Model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$
* Variable (in model) = Covariate = x
  - Treatment
  - Age
  - Gender
* Coefficient = Coef = $\beta \approx$ Association
* Y is called outcome, endpoint, but not CAUSAL

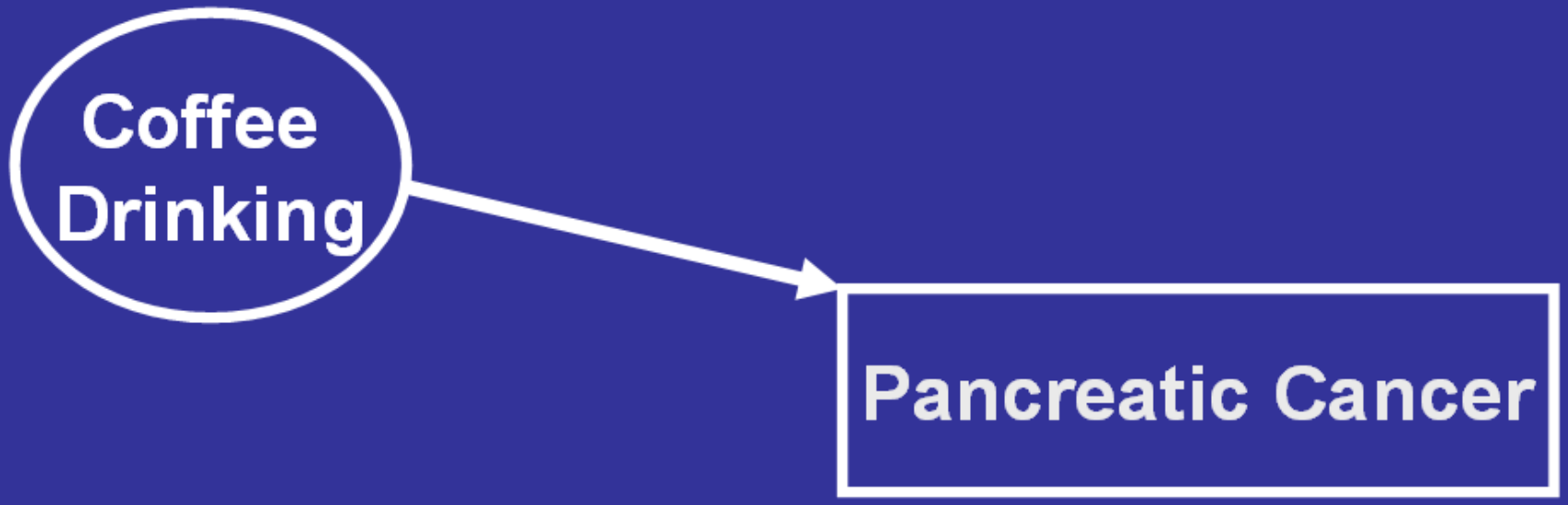
Covariates May Be

* Confounder(s)
* Effect Modifier(s)
* Other things

Confounding

* Two or more variables
* Known or *unknown* to the researchers
* Confounded when their effects on a common response variable or outcome are mixed together

Coffee and Pancreatic Cancer

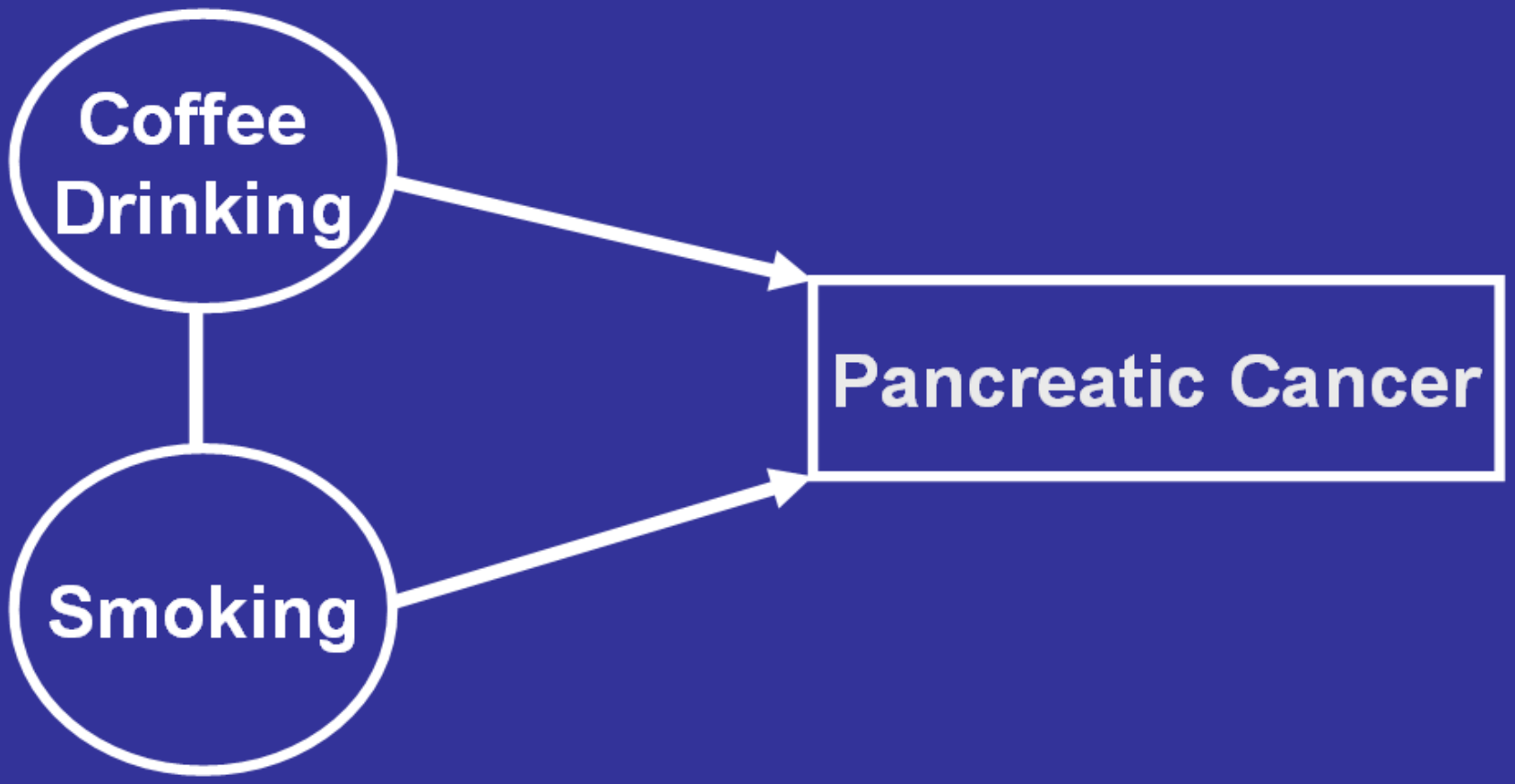


Coffee and Smoking

Confounding Example

* Relationship between coffee and pancreatic cancer, BUT
* Smoking is a known risk factor for pancreatic cancer
* Smoking is associated with coffee drinking but it is not a result of coffee drinking

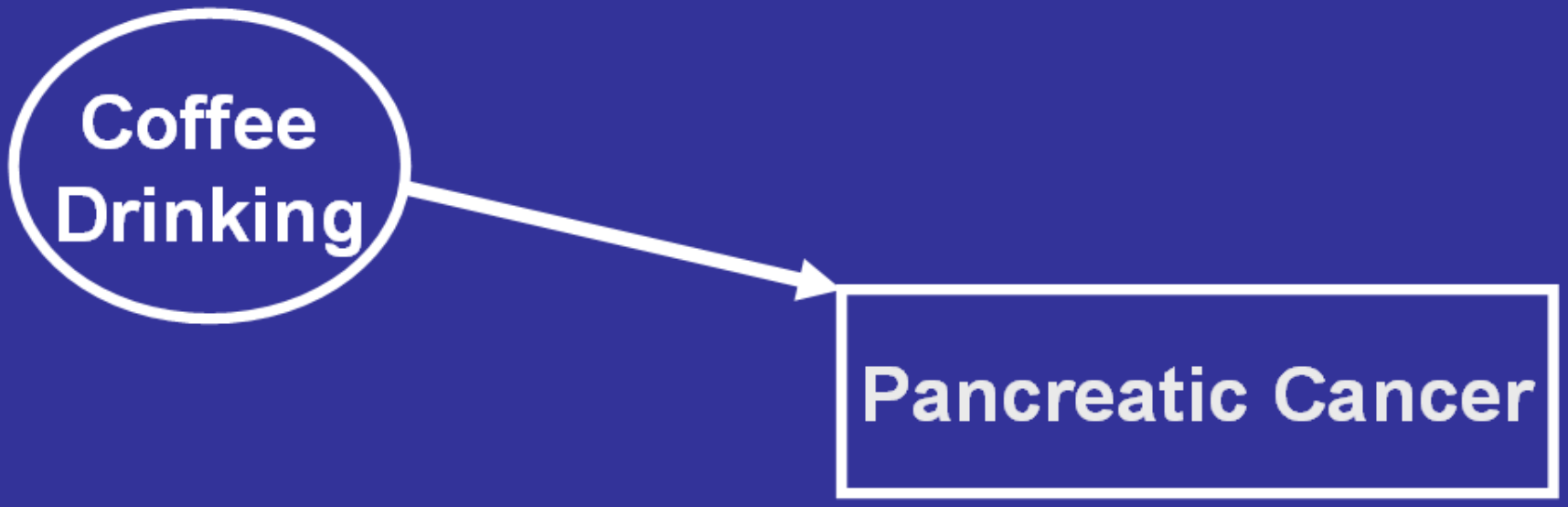
Coffee and Pancreatic Cancer



What is Confounding?

* If an association is observed between coffee drinking and pancreatic cancer
  - *Coffee actually causes pancreatic cancer, or*

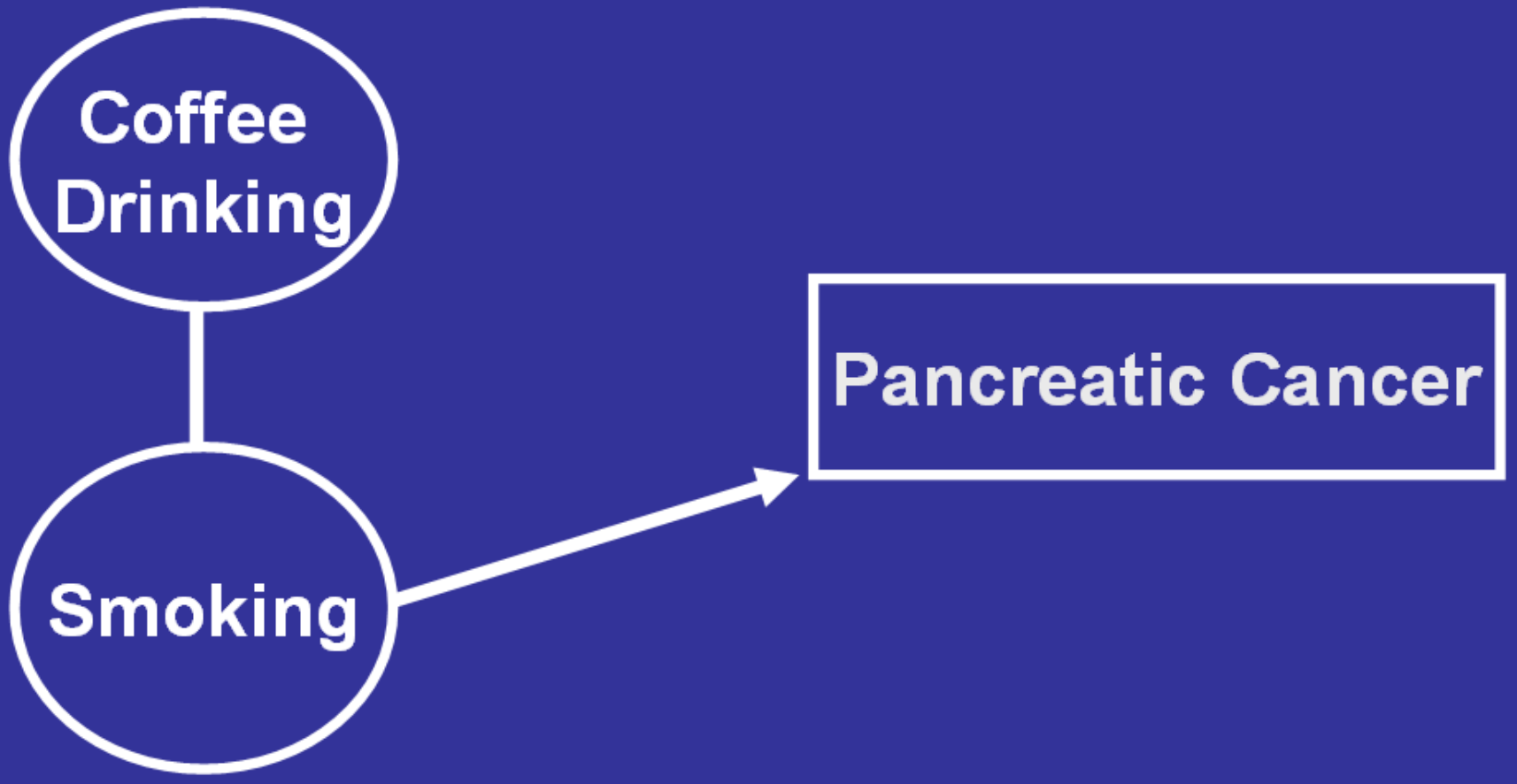
Coffee Causes Pancreatic Cancer



What is Confounding?

* If an association is observed between coffee drinking and pancreatic cancer
  - Coffee actually causes pancreatic cancer, or
  - *The coffee drinking and pancreatic cancer association is the result of confounding by cigarette smoking*

Smoking is a Confounder: Coffee does NOT cause Pancreatic CA



How to Handle Confounding

* ID potential confounders
  - MEASURE THEM!
  - In the data analysis use
    * Stratification, or
    * Adjustment (add the variable to the model)
* Fear the unknown!

Effect Modification

* Interaction
* Synergy
  - Could be larger or smaller
* The association between the outcome and another variable (e.g. the intervention) is modified by different levels of a third variable

Smoking, Asbestos Lung Cancer

* Smoking (alone) increases risk of lung cancer by A
* Asbestos exposure (alone) increases risk of lung cancer by B
* Smoking AND having asbestos exposure increases risk of lung cancer by MORE/LESS than A+B

Effect Modification

* www.massey.ac.nz/~wwwcphr/publications/introepi_teaching/ asite2_chapter%207.pdf

* The phrase effect modification, defined for different professions
  - Biostatisticians, public health workers, physicians, lawyers, biologists, epidemiologists,….

Prevalence, Incidence,….

* Prevalence
  - # with disease / # at risk
  - If you take a snap shot
  - How many diabetics in the US right now
  - Prevalence = Incidence * Duration
* Incidence
  - # NEW cases of disease (over a period of time) / # at risk during that period
  - How many new (incident) cases of diabetes diagnosed in 2007 / # who could develop dx

Sensitivity, Specificity

* Sensitivity: how good is a test at correctly IDing people who have disease
  - Can be 100% if you say everyone is ill (all have positive result)
  - Useless test with bad Specificity
* Specificity: how good is the test at correctly IDing people who are well

Bias

* Selection Bias
* Observational or interviewer bias

Selection Bias

* Prevalence Incidence bias
  - Exposed/impacted early? Might miss
    * Fatal episodes
    * Transient episodes
    * Silent cases
    * Case where evidence of exposure disappears with disease onset
* Non-respondent bias
  - Unwilling or unable to respond
  - Different exposures/outcomes from respondents?

Observational / Interviewer Bias

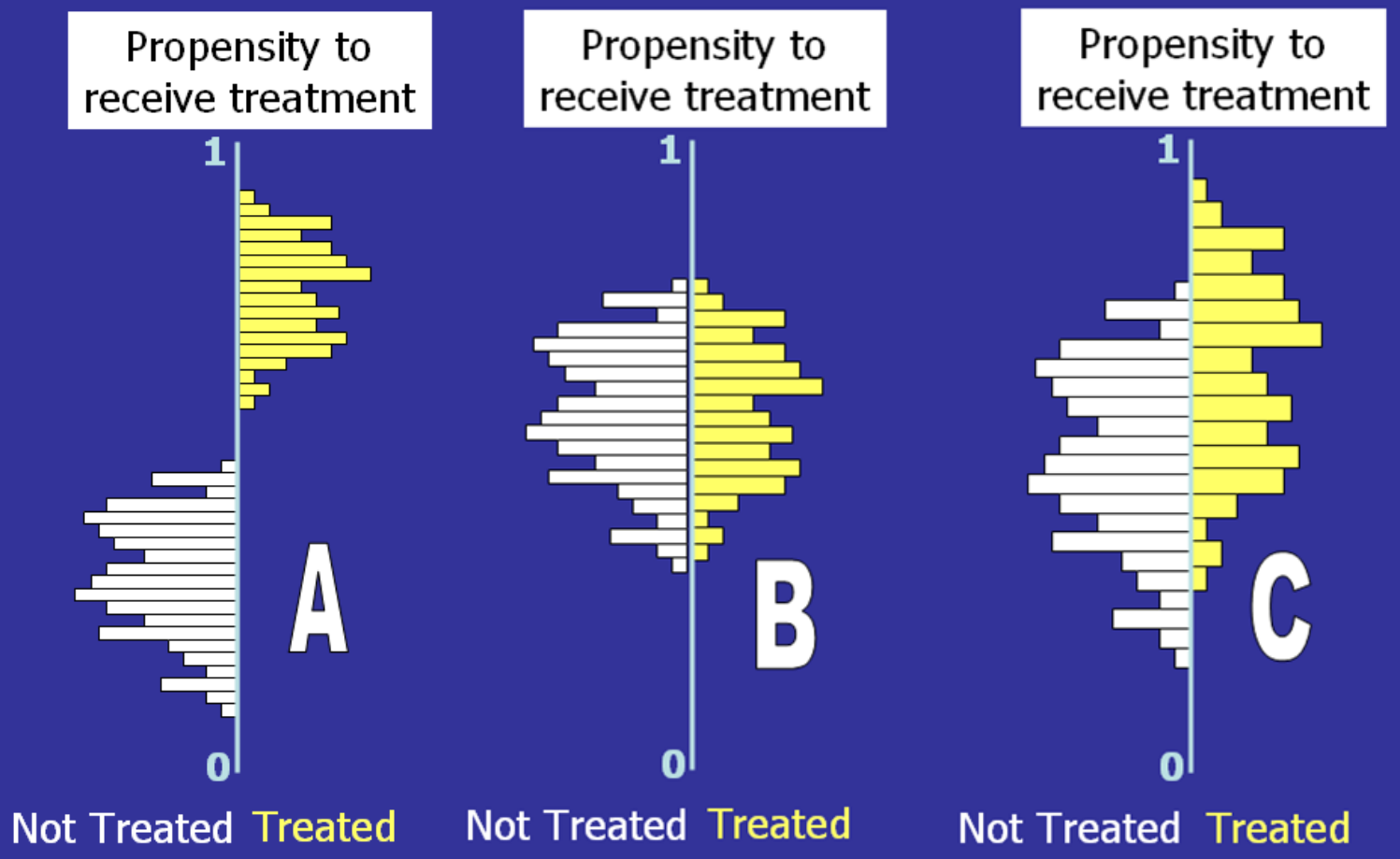
* Diagnostic suspicion bias
* Exposure suspicion bias
* Recall bias
* Family information bias

What do I do?

* Measure everything you can
* Build and investigate models
* Test those models on different data

* Try propensity scores
* Try other methods

How Much Overlap Do We Want?



Outline

- What is Epidemiology?
- Vocabulary (1)
* *Types of studies*
* Questions

Study Design Taxonomy

* Randomized vs. Non-Randomized
* Blinded/Masked or Not
  - Single-blind, Double blind, Unblinded
* Treatment vs. Observational
* Prospective vs. Retrospective
* Longitudinal vs. Cross-sectional

Ideal Study - Gold Standard

* Randomized
* Double blind / masked
* Treatment
* Prospective
* Parallel groups

Observational Studies

* *Case Reports*
* *Case Series*
- Cross-sectional Surveys
- Case-Control Study
- Cohort Study

Case Reports and Series

* Observations of patients with defined clinical characteristics
  - Certain disease
  - Cluster of symptoms
* Description of data without comparison groups
* Data from well defined group of people

Case Reports and Series

* Clear definitions of phenomenon
* Same definitions for all individuals in series
* Observations reliable and reproducible

* GOOD observational studies very useful

Case Reports and Series - Analyses

* Mean
* Standard deviation/error
* Proportions
* Confidence limits or intervals
* Separate data for subgroups
  - By sex, age, etc

Case Reports and Series

* Hypothesis formation
* Natural history
* Clinical experience
* Biased patient selection?
* Generalizability of results?
* Chance or characteristic?

Case Reports and Series

* Initial report of five cases of pneumocystis pneumonia in previously healthy, homosexual men
* CDC. Pneumocystis pneumonia-- Los Angeles. *MMWR* 1981; 30:250-2.

Observational Studies

- Case Reports
- Case Series
* *Cross-sectional Surveys*
- Case-Control Study
- Cohort Study

Crosssectional or Prevalence Surveys

* Observe prevalence and characteristics of disease
* Participant characteristics in a well defined population

Crosssectional or Prevalence Surveys

* Define population
* Derive a sample of the population
* Define the characteristics being studied
  - Standardized observations
  - Clearly defined
  - Methods of data collection applied equally to all study participants

Crosssectional or Prevalence Surveys

* Present
  - Prevalence (% or cases per 10 power of 5, etc)
  - Mean or median levels of relevant factors
  - Subset by important subgroups
* Analyses
  - Categorical
    * Chi-square, Fisher’s Exact Tests
  - Continuous
    * t-Test or other analyses

Observational Studies

* Cross-sectional
  - Collect a representative sample
  - Simultaneously classify by outcome and risk factor

		Outcome	
		Disease	No disease
Risk Factor	Y		
	N		

Crosssectional or Prevalence Surveys

* Descriptive
  - How common is the factor?
  - Characteristics of a group
  - Distribution of factors of interest (e.g. age)
* Associative
  - Relationships between factors
  - How do those with one factor differ from those without?

Crosssectional or Prevalence Surveys

* Inexpensive for common diseases
* More representative cases (vs. case series)
* Tend to be short (duration)
* Specific population
* Simultaneous wide variety of measurements

Crosssectional or Prevalence Surveys

* Unsuitable for rare diseases
* Unsuitable for disease of short duration
* High refusal rate - inaccurate prevalence estimates
* More expensive/time consuming than case control studies
* Time is the best/worst confounder of all

Crosssectional or Prevalence Surveys

* Prevalence data on overweight and obesity using measured height and weight in National Health and Nutrition Examination Survey (NHANES)
* Hedley AA, Ogden CL, Johnson CL, Carroll MD, Curtin LR, Flegal KM. Prevalence of overweight and obesity among US children, adolescents, and adults, 1999-2002. *JAMA* 2004;291:2847-50.

Observational Studies

- Case Reports
- Case Series
- Cross-sectional Surveys
* *Case-Control Study*
- Cohort Study

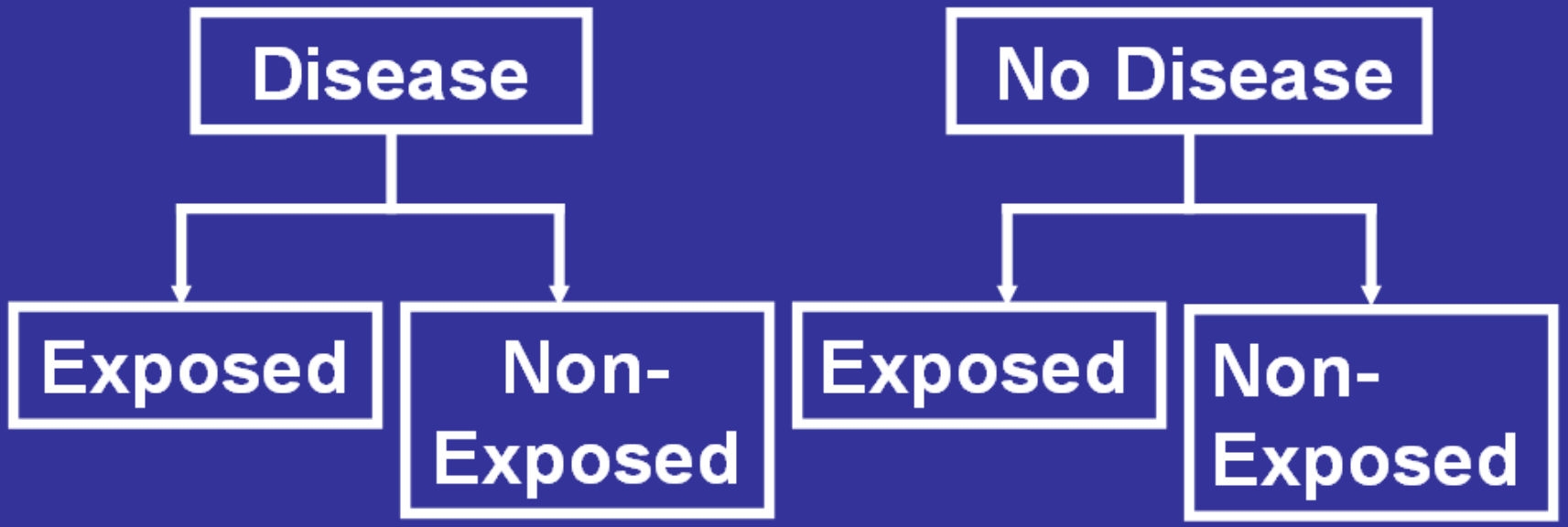
Case Control Studies

* Observations regarding possible associations between a disease and one or more hypothesized risk factors

Case Control Studies

* Compare the prevalence or level of the possible risk factor between
  - Representative group of disease subjects
    * CASES
  - Representative group of disease-free
    * CONTROLS
* Same population

Case Control Studies



Case Control Studies

* Cases represent all patients who develop disease
* Controls represent general ‘healthy’ population not developing the disease
* Information collected from cases and controls in the same way

Case Control Studies

* Standardized selection criteria from a well defined population
* Where?
  - Case registries
  - Admission records
  - Pathology logs
* High participation rate

Case Control Studies

* Perfect control group?
  - Next to never exists
* Standardized selection criteria from a well defined population
* Sample of
  - General population (gen pop)
  - Neighborhood
  - Families

Case Control Studies

* Cost to obtain controls?
* Multiple control groups!
  - Hospital control
  - Neighborhood control
* ‘Adjustment’ of results done during analysis (if subgroups large enough)

Case Control Studies

* Again,
  - All observations made using the same methods for cases and controls
  - Validity of measurement techniques established
* Selection, observation, and interviewer bias
* Use a 2x2 table

Case Control Studies

Character-istic/ Exposure	Presence of Disease		Total
	Number with Disease	Number without Disease	
Present	a	b	a + b
Absent	c	d	c + d
Total	a + c	b + d	N

Case Control Studies - Analyses

* Chi square or Fisher's exact tests
  - Proportion of cases exposed (a/a+c) compared to proportion of controls exposed (b/b+d)
* Continuous variables (especially in cross-sectional studies)
  - Mean levels of cases compared to controls or non-diseased subjects using Student's t test, non-parametric tests, etc.

Odds Ratio (OR)

* Odds are related to probability
  - Odds = p/(1-p)
* Probability of horse winning race is 50%, odds are 1/1
* Probability of horse winning race is 25%, odds are 1/3 for win or 3 to 1 <u>against</u> win

Odds

* If probability of diseased person being exposed is a/(a+c), odds are:

$$\frac{\frac{a}{a+c}}{1-\frac{a}{a+c}}$$

Odds and Odds Ratio

* Odds of exposure in cases: A/C
* Odds of exposure in controls: B/D

Odds Ratio (OR) = [A/C] / [B/D] = [AD] / [BC]

Relative Risk (RR)

* Risk in exposed [A/(A+B)] divided by risk in unexposed [C/(C+D)]
* But not used in case-control studies unless…..

Rare Disease, OR, RR

* A is small compared to B
  - All with exposure, # with disease vs. # without
* C is small compared to D
  - All without exposure, # w/ dx vs. # w/o dx
* Odds ratio estimates the relative risk well
  - OR is always further from unity
  - OR overestimates the magnitude of protective or harmful association

Case Control Studies

* Study the etiology of rare diseases
* Study multiple factors simultaneously
* Less time consuming and expensive
* ‘If assumptions are met’ associations and risk estimates are consistent with other types of studies

Case Control Studies

* Do not estimate incidence
* Do not estimate prevalence
* Relative Risk indirectly measured
* Bias is an issue
* Hard to study rare exposure
* Temporal relationship difficult to document

Case Control Studies

* Case-control design was able to identify relationship of exposure to stilbesterol during mother's pregnancy with occurrence of rare tumor in female offspring many years later
* Herbst AL, Ulfelder H, Poskaner DC. Adenocarcinoma of the vagina: Association of maternal stilbesterol therapy with tumor appearance in young women. *N Engl J Med* 1974;284:878-881.

Observational Studies

- Case Reports
- Case Series
- Cross-sectional Surveys
- Case-Control Study

* *Cohort Study*

Prospective or Longitudinal Cohort Studies

* Observations concerning associations between a given exposure (risk factor) and subsequent development of disease

Prospective or Longitudinal Cohort Studies

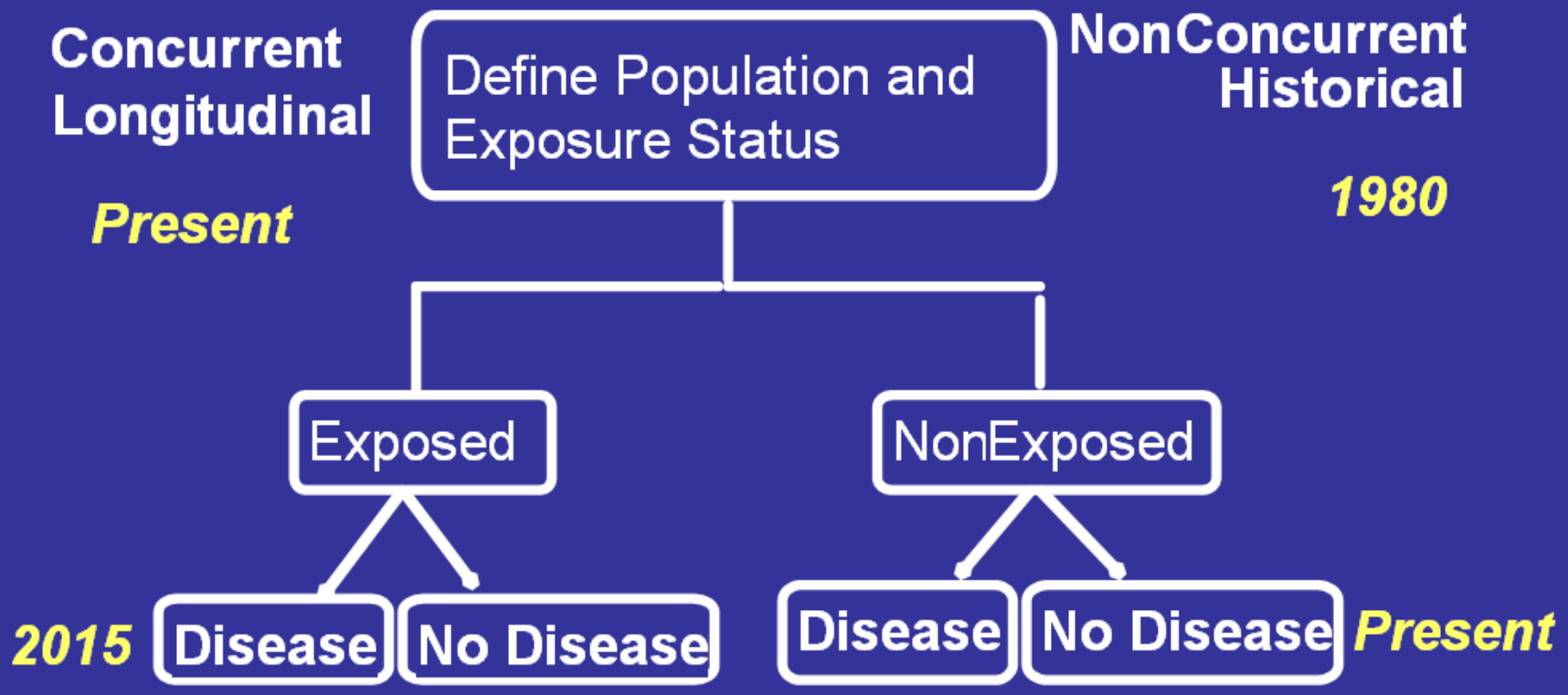
* Defined population is surveyed
* ID group with supposed risk factor
* ID similar group without risk factor
* Follow them forward in time
* Compare incidence rates between groups

* Could have a 0 in a cell on the 2x2 table

Prospective or Longitudinal Cohort Studies

* If non-concurrent prospective study
  - Really, retrospective or historical control study
* Defined population with presence/absence of exposure ascertained in accurate, object fashion in the past
  - Employment records
* Surveyed in present: disease occurrence
* Define incidence rates exposed/non

Prospective or Longitudinal Cohort Studies



Prospective or Longitudinal Cohort Studies

* Exposed and non-exposed are
  - Representative
  - Well-defined
* Absence of exposure
  - Well defined
  - Assumed maintained in non-exposed during the study

Prospective or Longitudinal Cohort Studies

* Outcomes (disease outcomes) well defined prior to study
  - Not changed during course of study
* Death – easy to define, ‘hard’ outcome
* Subjective symptoms – harder to define

Prospective or Longitudinal Cohort Studies

* Standard criteria applied to both exposed and non-exposed groups (again)
* Definitions of disease reliable and reproducible (again)
* Minimize loss to follow-up
  - Large non-response rates (>20%) raise questions as to the accuracy of the incidence rates

Prospective or Longitudinal Cohort Studies

* Calculate incidence for the study period in exposed, unexposed, and test using Chi square (x to 2nd power) or Fisher's exact test
* Measure association with relative risk (or odds ratio)
* 95% confidence limits (in a few weeks)
* Life-tables (last lecture)

Prospective or Longitudinal Cohort Studies

* More representative of cases than case-control (incident cases)
* More natural history information
* Incidence rates available
* Relative risk directly estimated

Prospective or Longitudinal Cohort Studies

* ‘Less’ bias
* Relationship to exposure
* Temporal relationship
* Rare exposure with frequent cases among exposed

Prospective or Longitudinal Cohort Studies

* LONG follow-up may be needed
* Free-living population follow-up is expensive
* Large population usually required
* Need baseline data
* Rare disease cannot be studied (rare exposure, yes)
* Bias (loss to follow-up, assessment, etc)

Prospective or Longitudinal Cohort Studies

* Prospective cohort study that showed early increase in risk of lung cancer and heart disease mortality and confirmed this over 50 years of follow-up
* Doll R, Hill AB. The mortality of doctors in relation to their smoking habits: A preliminary report. *Br Med J* 1954;228(i):1451-1455.
* Doll R, Peto R, Boreham J, Sutherland I. Mortality in relation to smoking: 50 years observations on male British doctors. *Br Med J* 2004;328:1519-1533.

Prospective or Longitudinal Cohort Studies

* Military medical records used to identify WW II head trauma exposure group and non-trauma comparison group who were traced and evaluated for dementia 50 years later

* Plassman BL, Havlik RJ, Steffens DC, et al. Documented head injury in early adulthood and risk of Alzheimer's disease and other dementias. *Neurology* 2000;55:1158-1166.

Summary

Observational Studies

* Case Reports/Case Series
* Cross-sectional Survey
  - NHIS (National Health Interview Survey)
* Case-Control Study
  - Groups with or without outcome
  - Determine who was exposed to risk factor
* Cohort Study
  - Follow a group for a while
  - Cardiovascular Health Study

Observational Studies are Useful

* May be only alternative
  - Smoking in humans
  - Long term HAART treatment
  - What happens in free living people (Cardiovascular Health Study)
* May be cheaper and faster than a trial

Do Not Always Agree

* Hormone Replacement Therapy
* Observational trials
* Women's Health Initiative (WHI)
* Publication bias?
* Incorrect analyses of observational studies?
* Different populations?

Questions?