Biostatistics for Health Care Researchers: A Short Course

Study Design

Presented by:

Patrick O. Monahan, PhD

Division of Biostatistics

Indiana University School of Medicine

Objectives

- Describe the elements of a research study
- Describe the main observational study designs
  - Epidemiologic
  - Outcomes
- Describe estimates of exposure-disease association, including relative, attributable, and population attributable risk and odds ratios.

How Does Medicine Advance

- Observation
- Informal Experimentation
- Formal Experimentation
  - Evidence based medicine (RCT)
  - Causal Inference

Elements of a Study

- QUESTION and TESTABLE HYPOTHESIS
- Outcome variable and its measurement
- Experimental Study Design
- Population
- Protocol
- Analyses
- Conclusion should answer question

Question and Hypothesis

- The QUESTION is the original problem that prompted the study
- The HYPOTHESIS is a rephrasing of the question in a statistically testable form
- Example
  - Question - does smoking cause lung cancer?
  - Hypothesis – Do people who smoke have a greater incidence of lung cancer than non-smokers?

Outcome Measure

- Reflects the hypothesis to be tested
- Clearly defined
- Appropriate for hypothesis (and question)

Epidemiologic Study Designs

- Case Series
- Assess association between risk factors and disease
- Ecologic
- Cross-Sectional
- Case-Control
- Prospective and Historical Prospective

Measures of Disease Frequency

- Prevalence (P) - proportion of individuals who have the disease at a specific time
- Cumulative Incidence (CI) - proportion of individuals who become diseased during a specified time interval
- Incidence Rate (I) - number of individuals who become diseased divided by person-time observed

Risk vs. Rate

- Prevalence - risk that an individual will be ill at a given point in time
- Cumulative Incidence - risk that an individual will develop the disease in a specific time interval
- Incidence Rate - instantaneous rate of development of disease in a population

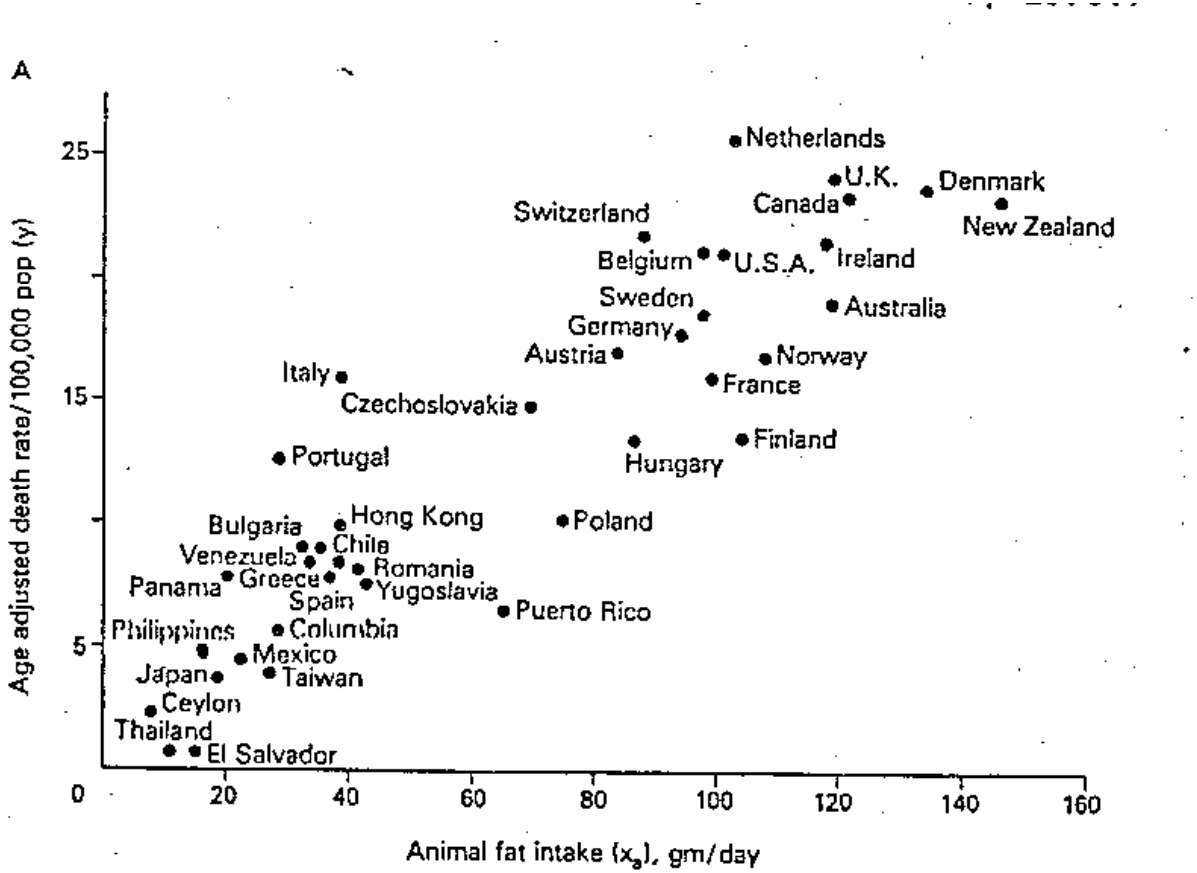
Ecologic Studies

- Purpose:
  - Describe patterns or trends on a geographic level
  - explore potential associations between community-level risk factors and disease
- Design:
  - Collect group level data (Country, state, city, etc.)
- Data:
  - Disease incidence or prevalence in each population
  - Risks and confounding factors of populations

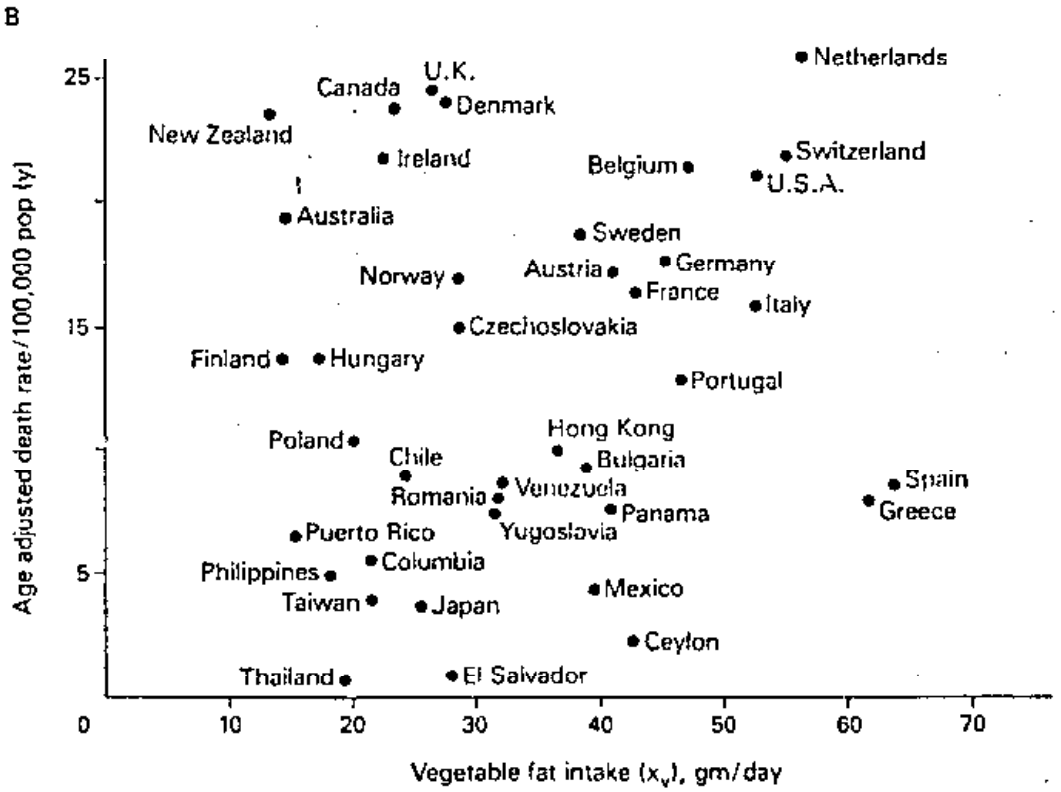
Ecological Study: Breast Cancer

- Data collected in the early 1970's from 39 different countries
- Relationship between per capita fat consumption and death rate from breast cancer
- Examined animal and vegetable fat separately

Fat Intake and Breast Cancer



Fat Intake and Breast Cancer



Ecologic Studies

- Advantages:
  - relatively quick and inexpensive
  - Allows estimation of effects not easily measurable for individuals
  - Permits exploratory analyses of potential factors in disease etiology
- Disadvantages:
  - Risk factors and disease endpoints are not measured on the same individuals
  - Difficult to control for confounders

Cross-Sectional Studies

- Purpose:
  - explore potential associations between individual-level characteristics and disease endpoints
- Design:
  - A single examination of a group of individuals
  - Physical Examination, Questionnaires
- Data:
  - Presence of disease (Prevalence)
  - Measurement of risk and confounding factors

Cross-Sectional Study: NHANES

- NHANES (National Health and Nutrition Examination Survey) Study
  - Random sample of the entire country
  - Physical and laboratory examinations
  - Dietary questionnaires
  - Looks at relationships between dietary intake and disease presence
  - Dietary data is self-report

Cross-Sectional Studies

- Advantages:
  - Disease and risk factor data are collected from the same individuals
  - Complete data collection
- Disadvantages:
  - Higher proportion of long term survivors
  - No data on time relationship between risk factors and disease development

Case-Control Studies

- Purpose:
  - Determine whether those with a disease, i.e., prevalent cases, differ from those without the disease
- Design:
  - Define and identify ‘cases’
  - Identify a comparison group - ‘controls’
  - Measure risk factors
- Data:
  - exposure to risk factors in case and controls

Measures of Association - Odds Ratios

		Disease: Yes	Disease: No
Exposure	Yes	P_1 A	$1 - P_1$ B
	No	P_2 C	$1 - P_2$ D

P_1 = prop. of exposed with disease = A/(A + B)

P_2 = prop. of non-exposed with disease = C/(C + D)

Odds ratio = OR = $\frac{P_1}{1-P_1} \div \frac{P_2}{1-P_2} = \frac{P_1(1-P_2)}{(1-P_1)P_2} = \frac{AD}{BC}$

Case-Control Study: Smoking

Cigarette Smoking	Lung Cancer: Cases	Lung Cancer: Controls
Yes	1350	1296
No	7	61

OR = (1350)(61)/(1296)(7) = 9.1

(Doll and Hill, Br. Med. J, 2:1271, 1952)

Case-Control Study: Febrile Seizures

	Crude OR	P-value for trend	OR adjusted for maternal age, education, and race	P-value for trend
Non-smoker	1.0		1.0	
Quit during pregnancy	1.6 (1.0-2.7)		1.2 (0.7-2.1)	
Smoked throughout pregnancy	2.1 (1.5-2.9)	<0.001	2.0 (1.3-2.8)	<0.001
1-10 cigs per day	2.1 (1.2-3.5)		1.6 (0.9-2.9)	
11-20 cigs per day	2.0 (1.2-3.5)		2.0 (1.2-3.4)	
=> 21 cigs per day	2.7 (1.2-6.1)	<0.001	2.6 (1.0-6.6)	<0.001

Numbers in parentheses are 95% confidence intervals

Study conducted in western Washington, 1983-1985.

Case-Control Studies

- Advantages:
  - Quick ascertainment of cases
  - Requires administration of questionnaires to a relatively small study population
  - useful for rare disease
- Disadvantages:
  - Potential for recall bias
  - Potential bias due to participation of non-representative group of controls (or cases)
  - generally can’t assess incidence or prevalence

Prospective Studies

- Purpose:
  - Establish incidence rates for disease
  - Estimate associations between risk factors and incidence of disease or survival
- Design:
  - Identify disease free cohort, follow over time
- Data:
  - collected repeatedly over time on presence of risk and confounding factors and disease development
  - Use physical exams and questionnaires

Prospective Study: British Male Doctors

- British Male Doctors (Doll and Peto, Br. Med. J, 2:1525, 1976):
- N=34,440
- Smokers: 140 lung cancer deaths per 100,000 subjects per year
- Non-smokers: 10 lung cancer deaths per 100,000 subjects per year

Measures of Association

- Relative Risk or Rate (RR) = CI_e/CI_u or I_e/I_u
- Attributable Risk or Rate (AR) = $CI_e - CI_u$ or $I_e - I_u$
- Population Attributable Risk (PAR) = AR $\times$ proportion exposed

British Male Doctors

- RR = 140/10 = 14. The risk of lung cancer death is 14 times higher in smokers than non-smokers
- AR = 140 - 10 = 130. The excess occurrence of lung cancer in smokers due to smoking is 130 per 100,000 subjects per year
- PAR = 130(.1)=13
  - if proportion exposed=.1

Prospective Studies

- Advantages:
  - Measures risk factors prior to the development of disease (i.e. time relationship)
  - Permits ascertainment of true incidence rates
  - useful for rare exposures
- Disadvantages:
  - Time
  - Expensive
  - Loss to follow-up bias
  - Difficult to study rare diseases

Historical Prospective Studies

- Purpose:
  - same as usual prospective study
- Design:
  - track disease occurrence between baseline and the present
- Data:
  - records for a cohort established some time in the past

Historical Prospective Study: Allegheny Co. Steelworkers

- All steelworkers employed for at least one month between 1/1/52 and 12/31/52
- Estimate exposure. Time employed in specific occupations (e.g., coke oven)
- Identify comparison populations, e.g. county workers, other mill workers

Allegheny Co. Steelworkers

Observed and Expected Mortality, 1953-1975, for Allegheny County Steelworkers

	Coke Plant N=3,558			Coke Oven N=2,041			Non-oven N=1,517		
Cause	Obs.	Exp.	RR	Obs.	Exp.	RR	Obs.	Exp.	RR
Respiratory Organs	121	83.4	*1.55	90	50.1	*2.03	31	32.4	0.95
Digestive Organs and Peritoneum	84	79.6	1.06	34	44.4	.74	50	34.6	*1.48
Genito-Urinary Organs	41	32.2	1.32	28	18.9	*1.56	13	12.8	1.02

*significant at .05 level

Association of Hepatitis C and Chronic Kidney Disease

- Using RMRS, identify subjects with Hep C test between 1994 and 2004 and no previous or concurrent CKD (N=8,224)
- CKD defined by elevated serum creatinine
- Use database to follow forward for CKD

Results

	CKD	No CKD
Hep C positive	428 (17.5)	2,049
Hep C negative	821 (14.3%)	4,926

$$\text{OR} = 428*4926 \,/821*2049 = 1.25$$

$$95\%\ \text{CI} = (1.10, 1.42)$$

$$\text{RR} = 1.28$$

Historical Prospective Studies

- Advantages:
  - Same as usual prospective studies, plus …
  - Data already exist, so study can be done almost as quickly as case-control studies
- Disadvantages:
  - limited data available
  - may be unrecorded changes in risk factors
  - few suitable cohorts exist, or have data available for study

Case Control vs. Cohort

- CASE CONTROL
  - Collect data on exposure presence in the past
  - Look back in time
  - Good for rare diseases
  - Small Samples
  - Inexpensive
  - Factors related to presence of disease
- COHORT
  - Collect data on disease development in future
  - Follow over time
  - Good for rare exposures
  - Large samples
  - Expensive
  - Factors related to development of disease

Outcomes Studies

- Purpose:
  - relate health care delivery to outcomes
- Design:
  - develop instrument to measure outcome if necessary
  - relate outcome to health care delivery (study design can be observational or experimental)

Common Outcomes Measures

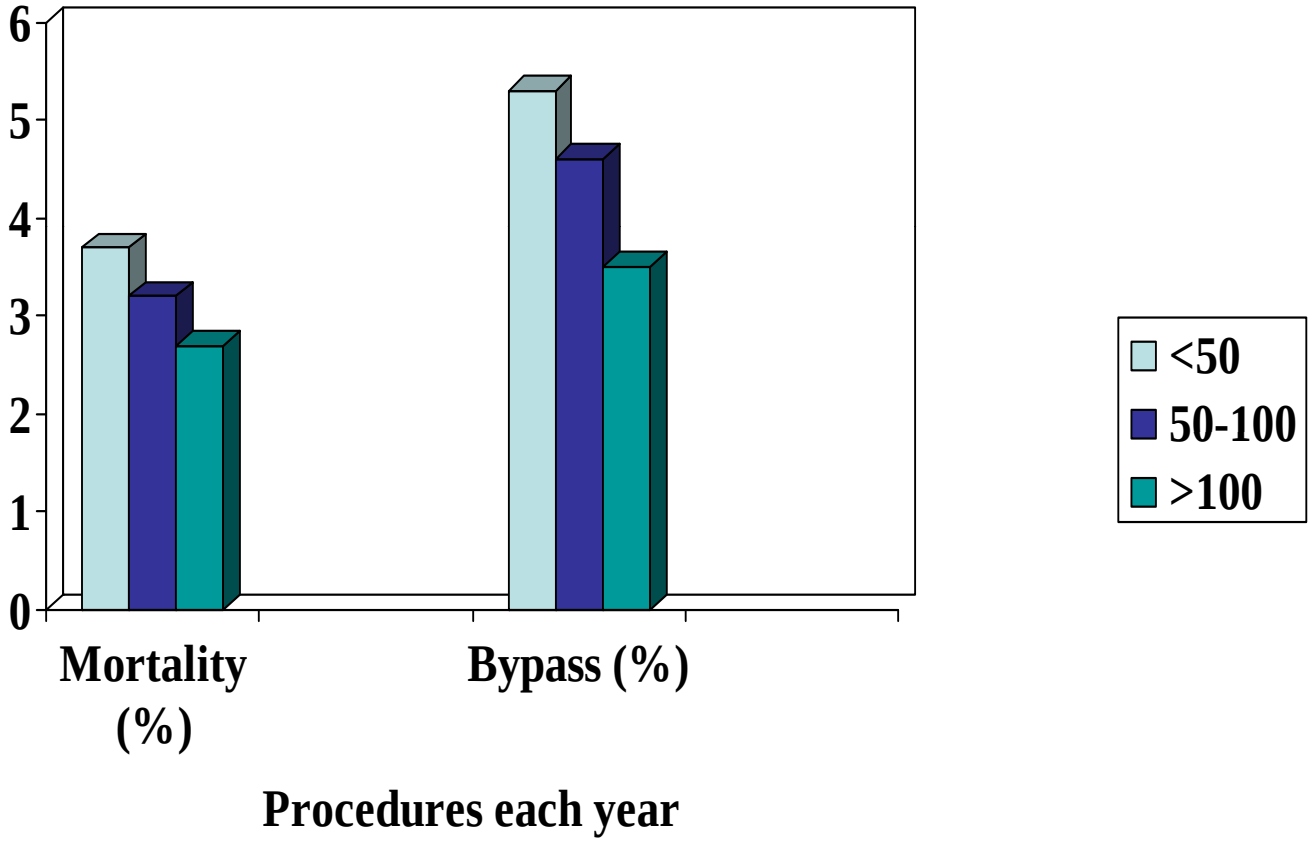
- Mortality
- Morbidity
- Disease Severity - PASI, NYHA
- Health Status - SF-36, SIP
- Quality of Life - QWB

Instruments are often specific to disease: e.g. AIMS

Health Care Delivery Studies

- Often Cross-sectional in design, using existing data bases
- Advantages:
  - Inexpensive
  - Questions not amenable to formal experiment
- Disadvantages:
  - Sparse and incomplete data
  - Population limitations
  - Treatment selection bias

Effect of Hospital Volume on Angioplasty Complications



Outcome Studies of Treatment

- BE VERY CAREFUL!
- Biased assignment to treatment
  - Treatments given to optimize outcomes
- Some statistical techniques available (e.g. propensity scores) but can only adjust for observed differences
- Cannot adjust for gross imbalance

Review

- Main observational study designs
  - Ecologic
  - Cross-sectional
  - Case-control
  - Cohort (prospective and historical)
- Estimates of exposure-disease association,
  - relative, attributable, and population attributable risk and odds-ratios