EPID600 (Spring 2007) module on Causal Inference

Objectives:

- Describe the elements of an epidemiologic study which must be considered before causality can be evaluated.
- Recognize the need for establishing causality in public health research.
- State the guidelines for judging whether an association is causal.
- Distinguish between real and spurious associations.
- Apply the guidelines in interpreting results of an epidemiologic study.
- Recognize how the presence or absence of an established causal relationship can enter into public health decision-making.
- Critically appraise a published journal article claiming to show epidemiologic evidence for a causal relationship.
- Apply the epidemiologic guidelines for causality to a research study to evaluate the degree to which these guidelines are satisfied by the authors' presentation.

Instructions:

1. **Read**: Aschengrau and Seage, ch. 15 - The Epidemiologic Approach to Causation . Answer the practice questions at the end of the chapter or at http://publichealth.jbpub.com/aschengrau/student_resources.cfm and check your answers (recommended, but optional) (animated flashcards, weblinks, and Powerpoint slides from the authors] can also be found at that URL)

2. We suggest that you first read the National Cancer Institute fact sheet on "Human Papillomaviruses and Cancer" at http://www.cancer.gov/cancertopics/factsheet/Risk/HPV/) for background.

3. Look over the case study questions and then read the case study reading: Schiffman MH, Bauer HM, Hoover RN, *et al.* Epidemiologic evidence showing that human papillomavirus infection causes most cervical intraepithelial neoplasia. *J Natl Cancer Inst* 1993; 85:958-964. (abstract. **UNC-CH**: full text)

4. (Optional, but earns credit) Before lab, submit the answers to the starred case study questions (numbers 5, 7, 8, and 9).

5. Read the lecture slides and attend the lecture (or read the speaker notes).

6. Work on the rest of the case study questions in **lab** and afterwards.

7. Agree on the answers, so the facilitator can submit the group's consensus answers by the following Sunday evening (EST).

Case Study Questions (NOTE: For some of these questions there may not be one "right answer".)

Preliminary Comments:

Cervical intraepithelial neoplasia (CIN) is considered to be a probable precursor of full blown invasive cancer of the cervix. By definition, CIN is limited to the epithelial lining of the cervix, the external entrance to the uterus. In the typical Pap smear, cells are obtained from the cervix, and these are examined under the microscope for evidence of atypical cells or for clearly abnormal cells classified into CIN-1, CIN-2, CIN-3, depending on the degree of abnormality. An additional procedure, namely a cervicovaginal lavage, is required to obtain specimens for HPV testing, as described in the journal article to be discussed.

NOTE: The relative risks presented in the paper are actually odds ratios serving as estimates of relative risks.

IMPORTANT! Hints for interpreting Table 2 in the HPV article by Schiffman.

- The authors present relative risks (RR) for the association of cervical cancer and several different risk factors (number of sexual partners, age at 1st intercourse, etc.) Three columns of RRs are presented. Each column has been adjusted for different confounding factors (read the notes listed below the table).

- For Example: RR#1 indicates that women with 10+ lifetime sexual partners have 4.4 times the risk of developing cancer as women who had 1 lifetime sexual partner. (The reference group is women with 1 lifetime sexual partner; therefore the RR for that group is 1.0).

- RR#2 indicates that adjusting for confounders does not change the RR for cancer for women with 10+ partners vs. 1 partner. RR#2 = RR#1 = 4.4.

- RR#3 is the RR for 10+ sex partners vs 1 partner, adjusting for age and HPV infection. This RR is 1.8, which indicates an 80% increase in risk associated with having 10+ partners, after accounting for the increase in risk associated with HPV infection.

1. a. Use the data in Table **1** to construct a 2 x 2 table and estimate the relative risk of CIN for women with Types 16 or 18 HPV.

b. Use the data in Table **4** to construct a 2 x 2 table and estimate the relative risk of **CIN 1** for women with Types 16 or 18 HPV.

2. A focused review of an epidemiologic study includes the following five aspects. Appraise the article on cervical intraepithelial neoplasia by Schiffman *et al.* in relation to these aspects.

a. Research hypotheses/questions: Are they clear? Are they relevant? Do they follow logically from what is already known, i.e. based on the existing literature?

b. Study design: Is it experimental or observational? What type of study is it? Is this design appropriate in light of past research, the research question and the nature of the disease and exposures?

c. Outcome variable: Is it relevant? How is it being defined/measured? How accurate is the outcome/disease measurement?

d. Exposure variable: Is it relevant? How is it being measured? With what level of accuracy? How is exposure quantified: how valid is the cutoff point for distinguishing exposed from unexposed? Are biological markers used to define exposure or is it self-report, medical records etc.?

e. Analysis: Does the analysis address the research question? Is the analysis appropriate for the study design and type of data collected? Do the analysis and presentation provide information on the precision of estimates?

3. Schiffman *et al.* made considerable efforts to review the original cytological diagnoses that serve as the basis for defining cases. What type of bias do these efforts address? What effect would these efforts, if they are successful, have on the estimated relative risk (see the authors' discussion on 962-1-1 [pg 962, col 1, para 1])?

4. The majority (n=319) of the 500 cases were defined as having condylomatous atypia, which the authors considered to be borderline rather than definite cases of CIN (see 959-2-5). Thus, it is possible that some of these borderline cases may have been misclassified as cases. Is it likely that the disease misclassification was nondifferential or differential with respect to the exposure, that is, HPV status? What is the likely effect of this misclassification on the estimate of relative risks given in Table 1?

**5. In the comparison of the RR#1 and RR#2 columns in Table 2, what potential bias are the authors addressing? What conclusion can you draw when you compare the RR figures from RR#1 and RR#2 for smoking?

6. Compare the RR differences in Table 2 for lifetime number of sex partners, between RR#1 and RR#3. Write a short (under 150 words) paragraph explaining the differences between these two columns.

**7. At the bottom of page 960, the authors describe two ancillary analyses based on subsets of the case group. What is the purpose of subsetting the data in this way? Do the results for these subsets increase your confidence in the validity of the overall conclusion? Why or why not?

**8. Schiffman *et al.* argue (in 961-1-2) that multivariate analyses including both lifetime numbers of sex partners and HPV test results pointed to HPV infection as the primary risk factor for each of the three categories of cases. What data in Table 4 support this argument? Explain.

**9. In their discussion section on page 962, the authors argue that the HPV association with CIN satisfies all of the accepted criteria for assessing causality. Which of these criteria are strongly satisfied and which somewhat weakly in the evidence discussed by Schiffman *et al.*?

10. a. Explain Schiffman *et al.*'s statement in 962-2-4 that even though some risk factors persist among HPV-negative women, this finding could result from errors in HPV measurements.

b. Suggest some ways in which HPV measurement errors could be reduced.

4/14/2004, 4/20/2006, 7/18/2006, 2/5/2007

ARTICLES

Epidemiologic Evidence Showing That Human Papillomavirus Infection Causes Most Cervical Intraepithelial Neoplasia

*Mark H. Schiffman, Heidi M. Bauer, Robert N. Hoover, Andrew G. Glass, Diane M. Cadell, Brenda B. Rush, David R. Scott, Mark E. Sherman, Robert J. Kurman, Sholom Wacholder, Cynthia K. Stanton, M. Michele Manos**

Background: **Experimental studies have provided strong evidence that human papillomavirus (HPV) is the long-sought venereal cause of cervical neoplasia, but the epidemiologic evidence has been inconsistent.** ***Purpose:*** **Given improvements in HPV testing that have revealed a strong link between sexual activity history and cervical HPV infection, we conducted a large case–control study of HPV and cervical intraepithelial neoplasia (CIN) to evaluate whether sexual behavior and the other established risk factors for CIN influence risk primarily via HPV infection.** ***Methods:*** **We studied 500 women with CIN and 500 control subjects receiving cytologic screening at Kaiser Permanente, a large prepaid health plan, in Portland, Ore. The established epidemiologic risk factors for CIN were assessed by telephone interview. We performed HPV testing of cervicovaginal lavage specimens by gene amplification using polymerase chain reaction with a consensus primer to target the L1 gene region of HPV. Unconditional logistic regression analysis was used to estimate relative risk of CIN and to adjust the epidemiologic associations for HPV test results to demonstrate whether the associations were mediated by HPV.** ***Results:*** **The case subjects demonstrated the typical epidemiologic profile of CIN: They had more sex partners, more cigarette smoking, earlier ages at first sexual intercourse, and lower socioeconomic status. Statistical adjustment for HPV infection substantially reduced the size of each of these case–control differences. Seventy-six percent of cases could be attributed to HPV infection; the results of cytologic review suggested that the true percentage was even higher. Once HPV infection was taken into account, an association of parity with risk of CIN was observed in both HPV-negative and HPV-positive women.** ***Conclusion:*** **The data show that the great majority of all grades of CIN can be attributed to HPV infection, particularly with the cancer-associated types of HPV.** ***Implications:*** **In light of this conclusion, the investigation of the natural history of HPV has preventive as well as etiologic importance. [J Natl Cancer Inst 85:958–964, 1993]**

The well-established association between sexual activity and the development of cervical neoplasia strongly implicates a sexually transmissible etiologic agent (*1,2*). Molecular studies have provided strong evidence that human papillomavirus (HPV) may be this agent (*3*), but the epidemiologic evidence has been weaker (*4,5*). HPV DNA is identified much more frequently in women with cervical neoplasia than in women with normal cervical cytologic diagnoses. Moreover, statistical adjustment for HPV infection has not explained the elevated risk of developing cervical neoplasia in women with multiple sex partners, suggesting that other venereally transmitted agents play an etiologic role (*6-9*). In addition, the estimated proportion of cervical neoplasia attributable to HPV infection in previous studies has been too low for one to conclude that HPV infection causes most cervical neoplasia.

Recently, improved HPV testing methods revealed for the first time a strong link between sexual activity history and cervical HPV infection (*10*). This finding prompted our large case–control study of cervical intraepithelial neoplasia (CIN) and HPV infection, which evaluates whether sexual behavior and the other established risk factors for cervical neoplasia influence risk primarily via HPV infection.

*See "Notes" section following "References."

Principles of Epidemiology for Public Health (EPID600)

Data analysis and causal inference

Victor J. Schoenbach, PhD home page

Department of Epidemiology
School of Public Health
University of North Carolina at Chapel Hill

www.unc.edu/epid600/

Abort, Retry, Fail

"Word for Windows 6.0: Self-Teaching Guide. . . . This book makes a good guide [but] surprisingly limits its audience to half by assuming that the reader is working in Windows."
– *ComputerUser*

[PC Magazine, 2/7/1995]

Data management

- Managing epidemiologic data is "mass production"
- A systematic, organized, professional approach is critical for detecting and avoiding problems

"You can never, never take anything for granted."

Noel Hinners, vice president for flight systems at Lockheed Martin Astronautics, whose engineering team reported measurements in English units that the Mars Climate Orbiter navigation team assumed were metric units.

Without the documentation, the data may be of little if any value (1995 NSFG)

```
0000000000000312222222402143041000
0000000000001144112131 070520310
0000000000003233112131 072331040
0000000000001116332227070350110
0000000000003133022221 02451121000
0000000000001111112131 02110041000
0000000000002111112131 07307131000
0000000000002122112131 01073041000
```


Data analysis and causal inference

- "Our data say nothing at all."
  (Epidemiology guru Sander Greenland, Congress of Epidemiology 2001, Toronto)
- Data are observer notes, respondent answers, biochemical measurements, contents of medical records, machine readable datasets, …
- What does one do with them?

Steps in data management

- Design the data collection process
- Write down all data collection procedures
- Train and supervise data collectors
- Monitor all data collection activities
- Document all data collection experiences
- Keep track of, document, and safeguard data

Data processing

- Review, edit, and code data forms, documenting exceptions and actions
- Convert to electronic form
- "Clean" data – check for illegal or improbable values, combinations of values
- Prepare summaries

Data exploration

- Examine the data – frequency distributions, cross-tabulations, scatterplots – be alert for surprises and suspicious findings
- Examine means and prevalence for factors of interest, overall and within interesting subgroups
- Look at associations, prevalence ratios, relative risks, odds ratios, correlations

Carry out focused data analysis

- Desirable to have a written analysis plan based on the research questions
- Typically carry out "crude" analyses and analyses controlling for important variables
- Methods of control: stratification, mathematical modeling

Stratified analysis

- Divide the dataset into subsets according to relevant covariables (e.g., age, sex, smoking, …)
- Examine the estimates and associations within each subset (unless there are too many)
- Take averages across the subsets

Mathematical modeling

- Express the outcome as some mathematical function of the relevant covariables
- "Fit" this function to the data, so that it models the relations in the data
- Interpret the resulting model to draw inferences about associations

Selecting a pattern to sew a pair of pants

- Want one that fits the need
- Can sew without a pattern, but takes time and may not look good
- Select a pattern that will be well received
  - Have you seen anyone wearing it?
  - Has it been featured in magazines

The strategy of statistical data analysis

Look for an available statistical model that will fit the situation (e.g., binomial, normal, chi-square, linear)

- Have others used it?
- Has it appeared in a methodology article?

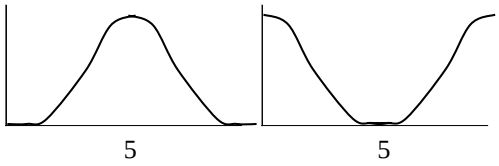
The strategy of statistical data analysis

Summarize the data in terms of the statistical model

- Mean
- Standard deviation
- Other parameters

But should always look at the data

- Distributions can have same mean and standard deviation but look very different – e.g., same mean:



Regression models - Conceptual

- Example:

Risk of CHD =

Age + BP + CHL + SMK

Regression models - Conceptual

- Suppose risk factors of:

Age	50 years
BP	130 mmHG systolic
CHL	220 mg/dL
SMK	30 pack-years

Regression models

- Risk of CHD = Age + BP + CHL + SMK
  Age = Years x risk increase per year
  BP = mmHG x risk increase per mmHG
  CHL = mg/dL x risk increase per mg/dL
  SMK = pack-years x risk increase per pack-year

Regression models

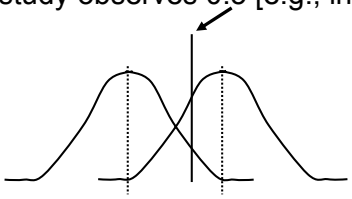
- Risk = $\beta_0 + \beta_1$Age + β_2BP + β_3CHL + β_4SMK
  β_0 = baseline risk
  β_1 = risk increase per year
  β_2 = risk increase per mmHG
  β_3 = risk increase per mg/dL
  β_4 = risk increase per pack-year
- Use the data and statistical techniques to estimate β_1, β_2, β_3, β_4.

P-values and Power

- P-value: "the probability of obtaining an interesting-looking sample from a boring population" (1 – specificity)
- Power: "the probability of obtaining an interesting-looking sample from an interesting population" (sensitivity)

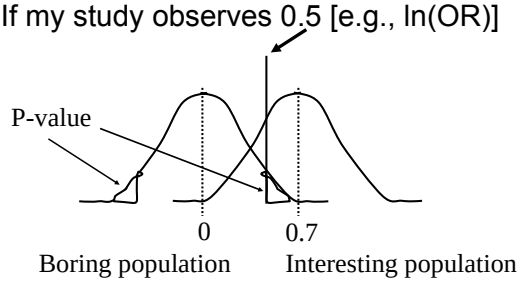
The P-value

If my study observes 0.5 [e.g., ln(OR)]



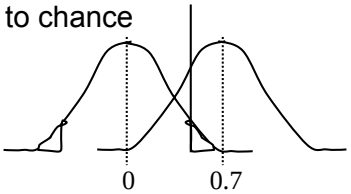
The P-value

If my study observes 0.5 [e.g., ln(OR)]

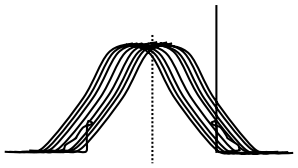


The Problem with P-values

But the P-value does <u>not</u> tell me the probability that what I observed was due to chance

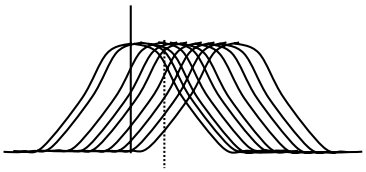


If I study only boring populations



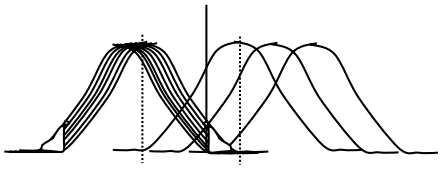
Boring populations

If I study only interesting populations



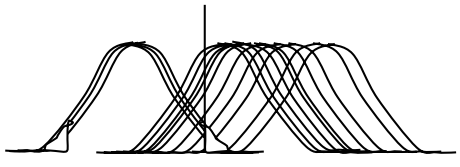
Interesting populations

Many boring populations



Boring populations Interesting populations

Many interesting populations



Boring populations Interesting populations

Do I study boring populations?

That probability depends on how many boring populations there are. If I study

10 interesting populations

100 boring populations

We expect me to obtain 9 interesting samples from the interesting populations and 5 from the boring populations

P-values and predictive values

Results:

14 interesting samples

5 came from boring populations

Probability that an interesting sample came from a boring population:

5/14 = 36% – not 5%!

Analogous to positive predictive value

P-values and predictive values

Samples	Populations: Interesting	Boring	Total
Interesting ("positive")	9	5	14
Boring ("negative")	1	95	96
Total	10 (cases)	100 ("noncases")	110

What should guide data analysis

- What are the research questions?
  - Estimate means (e.g., cholesterol) and prevalences (e.g., HIV)
  - Assess associations (e.g., Is blood lead level associated with elevated blood pressure?; Do prepaid health plans provide more preventative care?)

Trend analysis

Causal relations and public health

Many public health questions hinge on causal relations, e.g.

- Does dietary fiber prevent colon cancer?
- Do abstinence-only sex education programs raise the age of sexual debut?
- What level of arsenic in drinking water is harmful?
- Does higher patient volume reduce knee replacement complication rates?

Conceptual issues in causal relations

- In general we cannot "see" causal relations but must infer their existence.
- "Proving" causation means creating a belief – our own and others'.
- Causal inference is therefore a social process.
- What we regard as "causes" depends on our conceptual framework.

Pre-20th century causal discoveries

- Food poisoning from shellfish, pork
- Plumbism from wine kept in lead-glazed pottery (Romans)
- Contagion (isolation, quarantine)
- Scurvy and citrus fruit (James Lind)
- Scrotal cancer in chimney sweeps (Percival Pott)

Pre-20th century causal discoveries

- Smallpox vaccination
- Cowpox vaccination (Edwin Jenner)
- Waterborne transmission of typhoid fever (William Budd) and cholera (John Snow)
- Person-to-person transmission of measles (Peter Panum)
- Puerperal fever and handwashing (Ignaz Semmelweis)

Rise of the germ theory

- Invention of the microscope led to the science of bacteriology
- Laboratory experiments provided powerful evidence
- Henle-Koch postulates adopted for proving that a microorganism causes a disease

Henle-Koch postulates

1. The parasite must be present in all who have the disease;
2. The parasite can never occur in healthy persons;
3. The parasite can be isolated, cultured and capable of passing the disease to others

E.H. Carr – *What is history*?

"History … is 'a selective system' … of causal orientations to reality.… from the infinite ocean of facts [and] … the multiplicity of sequences of cause and effect [the historian] extracts those, and only those, which are historically significant; and the standard of historical significance is his ability to fit them into his pattern of rational explanation and interpretation. Other sequences of cause and

E.H. Carr – *What is history*?

effect have to be rejected as accidental, not because the relation between cause and effect is different, but because the sequence itself is irrelevant. The historian can do nothing with it; it is not amenable to rational interpretation, and has no meaning either for the past or the present." (E.H. Carr, *What is History*, p. 138).

When to act?

"All scientific work is incomplete – whether it be observational or experimental. All scientific work is liable to be upset or modified by advancing knowledge. That does not confer upon us a freedom to ignore the knowledge we already have, or to postpone the action that it appears to demand at a given time."

A.B. Hill, The environment and causation, p. 300

Is cigarette smoking harmful to health?

- Surgeon General's Advisory Committee on Smoking and Health, chaired by Dr. Luther Terry.

Surgeon General's Advisory Committee on Smoking and Health

- Long existing concern about health effects of smoking
- Accumulation of scientific studies in 1950's
- Committee of the Royal College of Physicians in Britain issued a report in 1962 indicting cigarette smoking as a cause of lung cancer and bronchitis and probably of CVD
- Major health problem, major industry, $$$

"Criteria for causal inference"

1. Strength of the association
2. Consistency - replication
3. Specificity of the association
4. Temporality
5. Biological gradient
6. Plausibility
7. Coherence
8. Experiment
9. Analogy

1. Strength of the association

- Is there an association?
- Is there really an association? (not chance, not bias, not confounding)
- Stronger associations less likely to be entirely due to confounding
- How strong is strong?

How strong is strong?

Relative risk	"Meaning"
1.1-1.3	"Weak"
1.4-1.7	"Modest"
1.8-3.0	"Moderate"
3-8	"Strong"
8-16	"Very strong"
16-40	"Dramatic"
40+	"Overwhelming"

2. Consistency - replication

- Has this association been observed in other studies?
- By other investigators?
- Working independently?
- With different methods?
- (Problematic for one-time events)

3. Specificity of the association

- Does what we see conform to what our conceptual model says we should see?
- If we expect a specific causal relation, is that what we see?
- The more accurately we define the factors, the greater the relative risk.

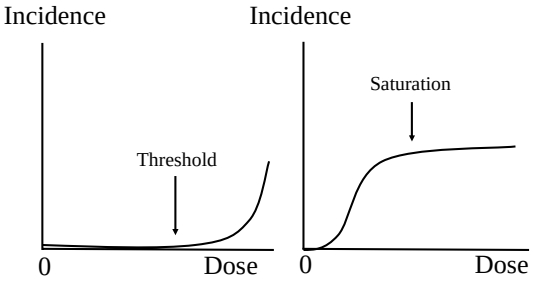
4. Temporality

- In everyday life, a cause must be present before its effects, at least by an instant.
- Subclinical disease states may be present long before the outcome is detected.

5. Biological gradient

- "Dose-response" relation – if we expect one.
- Often think that bias would not produce a dose-response relation.
- Biological model might predict threshold and/or saturation.

Possible dose-response curves



6. Plausibility

- Can we explain the relation on the basis of existing biological (psychological, social, etc.) knowledge?
- Problematic for new types of causes

7. Coherence

Does all of what we know fit into a coherent picture?

- Descriptive epidemiology of the exposure and disease by person, place, and time
- Related biological, economic, geographical factors

8. Experiment

Epidemiologic experiments can provide unique evidence – exposure precedes outcome; substitute population may be valid.

- –Randomized trials
- –Quasi-experimental studies
- –Natural experiments

9. Analogy

- Like plausibility, but weaker
- We are readier to accept something similar to what we've seen in other contexts.
- This criterion illustrates the point that causal inference involves getting people to change their beliefs

Causal inference in epidemiology and law

- Decision about facts must be reached on the evidence available
- Emphasis on integrity of the **process** of gathering and presenting information
- Requirement for adequate representation of contending views

Epidemiology and the legal process

- Use of standards of certainty for various potential consequences.
- Reliance on procedural (methodological) safeguards, since facts are established only as findings of an investigatory process.
- Justice (i.e., proper procedures / methodology) must be done and also <u>seen</u> to be done

Epidemiologic decision-making and the legal process

- Increasingly, epidemiologists and epidemiologic data are entering the courtroom.
- E.g.'s, Benedectin, silicon breast implants, environmental tobacco smoke, diesel exhaust.