EPID600 (Spring 2007) module on Selection Bias

Objectives:

- Identify the major sources of error in epidemiological studies.
- Differentiate between internal and external validity.
- Discuss sources of variability and threats to validity in published or planned epidemiologic research studies.
- Define selection bias and differentiate it from external validity.
- Give examples of selective forces on study populations and how they can be reduced or eliminated from epidemiological studies.
- Differentiate among the external, target, actual, and study populations and explain how selection bias can be conceptualized in relation to them.
- Calculate selection probabilities from hypothetical data and draw implications for the presence of selection bias in those data.

Instructions:

1. **Read**: Aschengrau and Seage, ch. 10 - Bias . 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. Look over the case study questions and then read the case study reading: Peterman TA. (1995). Can we get people to participate in a study of sexual behavior? *Sexually Transmitted Diseases* 22, 164-168. (**UNC-CH**: full text))
   and
   J. Frantzen, T.G.W. Speel, L.A. Kiemeney and E.J.H. Meuleman. Cardiovascular Risk Among Men Seeking Help for Erectile Dysfunction. *Annals of Epidemiology* 2006 (February); 16(2):85-90. (abstract, **UNC-CH:** full text)

3. (Optional, but earns credit) Before lab, submit the answers to the starred case study questions (numbers 1, 3, 4, 8, and 12).

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

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

6. 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".)

Erectile dysfunction and cardiovascular disease: The following questions refer to the article "Cardiovascular Risk Among Men Seeking Help for Erectile Dysfunction", by J. Frantzen, T.G.W. Speel, L.A. Kiemeney and E.J.H. Meuleman. *Annals of Epidemiology* 2006 (February);16(2):85-90.

**1. “Erectile dysfunction (ED) is a multifactorial disease of the aging male affecting millions of men worldwide. In the Netherlands, on average 13% of men aged 40 years and older are affected (1).” (p85,c1) If we interpret the phrase “aged 40 years and older” as 40-79 years in 2000, about how many men in this age group in the Netherlands in 2000 would have been affected? Include the calculation. Hint: remember the Census Bureau's International Data Base.

2. “The prevalence increases with age: 6% of men aged 40-49 years, compared to 38% of men aged 70-79 years.” (p85,c1) Suppose that these age-specific prevalences and the overall prevalence of 13% were true for men aged 40-79 years in the Netherlands in the year 2000. If the prevalence among men age 50-59 years old were 8%, what would the prevalence of ED have been for men aged 60-69 years? Include the calculation.

**3. Patient A (hypothetical) was born on January 1, 1930, entered the study on January 1, 1996 with 10 years of prior medical information, was first diagnosed with ED January 1, 2000 and with cardiovascular disease on January 1, 2001. He died on March 1, 2001. Give all answers in the number of months, rounded to the nearest month.

3a. How much follow-up time did patient A contribute for the calculation of ED incidence during the period before introduction of sildenafil?

3b. How much follow-up time did patient A contribute for the calculation of ED incidence during the period after introduction of sildenafil?

3c. How much follow-up time did patient A contribute for the calculation of cardiovascular disease incidence after his diagnosis with ED?

**4. The study design being employed to study the incidence of CVD is a . . . (Choose one best answer.) (8 hpts) Submission instruction: please enter only the question number and the letter for your choice.

A. Cross-sectional study
B. Case-control study with incident cases
C. Cohort study
D. Ecologic study
E. Intervention trial

5. “Prevalence of CVD was higher among men with ED compared to controls. The odds ratio was 2.07 [95%-CI 1.67-2.56] for the period before the introduction of sildenafil.” (p86c2) Suppose that this odds ratio had been calculated from a 2 x 2 table in which the prevalence of CVD among the controls were 7%. Estimate the prevalence in the men with ED during the period before the introduction of sildenafil (show the calculation). Why do we know that it must be less than 14.5% even without calculating it?

6. Which one of the following numbers gives the best estimate of the incidence proportion (cumulative incidence) of cardiovascular disease from the data in Figure 1? Show the calculation.

A. 0.0212 B. 0.0221 C. 0.0244 D. 0.0250 E. 0.0818

7. ""Depending on age category, the number of men was 1.5 to 2.1 as high in the period after compared to the period before.” Use data from the article to show the calculation of the 1.5 in the quoted sentence.

**8. Table 2 presents estimates of CVD incidence for men with and without ED in the periods before and after the introduction of oral sildenafil. Without regard to the confidence intervals, which one of the following estimates is the least meaningful or interpretable? Briefly support your answer. (Choose one best answer and provide a brief supporting statement.)

A. 50.8 per 1000 person-years.
B. 29.4 per 1000 person-years.
C. 34.3 per 1000 person-years.
D. 24.9 per 1000 person-years.
E. 23.6 per 1000 person-years.
F. 23.9 per 1000 person-years.

9. “The estimation of the risk of incident CVD was more precise for the period after than the period before the introduction of sildenafil (Table 2).” (p86c2). Which of the following risk estimates in Table 2 was most precisely estimated. Support your answer by stating the most relevant number or statistic from the table. (60 words maximum)

A. ED subjects before introduction of sildenafil (period A)
B. Control subjects before introduction of sildenafil (period A)
C. ED subjects after introduction of sildenafil (period B)
D. Control subjects after introduction of sildenafil (period B)

10. “The relative risk [for Fig. 2B] was estimated at 1.7 [95%-CI 0.9-3.3] using the proportional hazards model.” (p86c2). Calculate the corresponding relative risk from the data in Table 2. Show the calculation to three significant figures.

11. “We know that about a quarter of men suffering from ED did consult a physician before the introduction of sildenafil (1). . . . Erectile dysfunction might, therefore, be relatively underdiagnosed in the period before the introduction.” (p87c2-p88c1). Suppose that this estimate of one-quarter is accurate. Suppose also that the sensitivity of physician diagnosis (and therefore of being counted as a case by Frantzen *et al.*) when a man complains about ED is 90%. Assuming 100% specificity of physician diagnosis, what true underlying incidence rate of ED for the 40,388 Netherlands men in the upper half of Table 1 is reflected in the observed rate of 5.3 per 1000 person-years?

**12. The substantial underreporting of ED by men and underdiagnosis by general practitioners create significant potential for selection bias. Conceivably, selection bias was responsible for the observed association between incident ED and prevalent CVD during the period before sildenafil. Alternatively, such an association might actually exist, but it might not have been observed due to selection bias following the substantial increase in the number of men consulting a general practitioner for ED following the introduction of sildenafil. Conceivably selection bias could also have influenced the appearance of elevated CVD risk in men whose ED was noted before the introduction of sildenafil or the absence of such an elevation after the introduction of sildenafil. Describe a scenario in which selection bias would have occurred and had one of these influences. Do you think that any of these influences occurred?

13. For group discussion in classroom; optional for Internet: “The overall prevalence of cardiovascular diseases was about 8% in both, the period before and after the introduction of sildenafil. Prevalence of CVD was higher among men with ED compared to controls. The odds ratio was 2.07 [95%-CI 1.67-2.56] for the period before the introduction of sildenafil. After the introduction, the odds ratio was significantly lower, namely, 1.38 [95%-CI 1.21-1.57].” These confidence intervals are apparently incorrect. The 95% confidence interval for the odds ratio in a 2x2 table is obtained by first estimating the 95% confidence limits for the natural logarithm of the odds ratio and then exponentiating (taking anti-logarithms). The lower 95% confidence limit is ln(OR) - 1.96 x s.e.[ln(OR)] ; the upper 95% limit is ln(OR) - 1.96 x s.e.[ln(OR)], where s.e.[ln(OR)] is the standard error of the ln(OR) estimate. The standard error is obtained as the square root of its estimated variance, 1/a + 1/b + 1/c + 1/d, where a, b, c, d are the cells of the familiar 2x2 table. Use the information in the quoted paragraph, the first column of Table 2, the number of prevalent cases of CVD from Figure 1, and the 8% CVD prevalence for both time periods to show that the authors most likely omitted the multiplier of 1.96 in their calculation of the standard errors. Without intending to, they calculated 68% confidence intervals!

Can We Get People to Participate in a Study of Sexual Behavior?

THOMAS A. PETERMAN, MD, MSC

Background and Objectives: **Interpreting the results of any study requires careful analysis to determine how selective participation may have influenced the findings. Participation in a study of sexual behavior may be influenced by who is doing the study, who the participants are, what it means to participate, what the study involves, and what is meant by sexual behavior.**

Goal of this Study: **Identify potential sources of bias that could be detected by asking the question "Can we get people to participate in a study of sexual behavior?"**

Study Design: **Review of important examples of participation bias from the literature. Examination of a convenience sample of key studies of sexual behavior to see how they have addressed participation-related issues.**

Results: **There are many examples of studies where misleading results were caused by participation bias. Many key studies of sexual behavior did not address fully the potential impact of selective participation.**

Conclusion: **No study is completely without participation bias. Understanding potential sources of bias is essential when designing a study or interpreting the results.**

DECIDING WHETHER OR NOT we can get people to participate in a study of sexual behavior seems relatively simple. However, some of the persons who are eligible for a study may not enter the study, and some who do enter will not accurately complete the entire study. Therefore, interpreting the results of a study requires detailed analysis of how selective participation may have influenced the findings. For example, a *Time* cover story reported that 70% of women who had been married for at least 5 years were having extramarital affairs.[1] However, this result was based on a postal survey that had a 4.5% response rate. The results of every study will be influenced by who is doing the study, who the participants are, what it means to participate, what the study involves, and what is meant by sexual behavior.

Address correspondence to Dr. Thomas Peterman, Mailstop E02, Centers for Disease Control, Atlanta, GA 30333.

Reprint requests: Technical Information Services, Mailstop E08, Centers for Disease Control, Atlanta, GA 30333.

From the Division of STD/HIV Prevention, National Center for Prevention Services, Centers for Disease Control and Prevention, Atlanta, Georgia

This paper discusses the potential impact of these components of the question "Can we get people to participate in a study of sexual behavior?" Next, it reviews some important studies of sexual behavior to see how the investigators addressed participation problems in their study.

We

The results of a study may be influenced profoundly by who is doing it—who "we" are.[2,3] There are many examples of this. Officials of the tobacco institute and the Environmental Protection Agency point out each other's potential for bias in interpreting data on secondhand smoke. Many consumers recognize this general potential for bias, so advertisers are careful to point out that the surveys they quote were conducted by "an independent research organization." Some medical journals now require authors to list affiliations and sources of funding that may represent a conflict of interest.

The research team also can influence the outcome of a study before the data are analyzed. The organization sponsoring the research may influence participation rates or the type of study that can be done. Participation in a U.S. government study is likely to be different from participation in a study sponsored by a community activist group such as ACT UP. The organization also might influence the general tone of the responses.

The staff conducting the study may influence initial agreement to participate but probably has a greater influence on making participants comfortable enough to respond to individual questions. An individual may be more (or perhaps less) candid when talking to a person perceived as a peer compared with a university re-

ELSEVIER

Cardiovascular Risk Among Men Seeking Help for Erectile Dysfunction

J. FRANTZEN, PHD, T.G.W. SPEEL, MD, L.A. KIEMENEY, PHD,
AND E.J.H. MEULEMAN, MD, PHD

PURPOSE: The introduction of sildenafil put the risk of cardiovascular disease (CVD) among men with erectile dysfunction (ED) on the agenda of physicians. The question arose, Is EDsentinel to CVD? We sought to answer this question in the present study.
METHODS: A historical cohort study was set up using medical records of general practices all over the Netherlands. Incident cases of ED were selected before and after the introduction of sildenafil using a catchment population of 60,000 men aged 35 to 74 years. Two to three men without ED (controls) were, subsequently, matched to each case. Incidence of CVD was determined for cases and controls, respectively.
RESULTS: Overall, incidence of ED doubled from 5.3 per 1000 men-years in the period before introduction of sildenafil to 10.1 after the introduction. The relative risk of incident CVD among men with ED compared to controls was 1.7 [95%-CI 0.9–3.3] before the introduction and 1.1 [95%-CI 0.6–1.8] afterwards.
CONCLUSIONS: While ED could be seen as a marker for CVD before the introduction of sildenafil, it was clearly not afterwards.
Ann Epidemiol 2006;16:85–90.

KEY WORDS: Erectile Dysfunction, Cardiovascular Disease, Sildenafil, Prevention.

INTRODUCTION

Erectile dysfunction (ED) is a multifactorial disease of the aging male affecting millions of men worldwide. In the Netherlands, on average 13% of men aged 40 years and older are affected (1). The prevalence increases with age: 6% of men aged 40–49 years, compared to 38% of men aged 70–79 years.

The introduction of oral sildenafil was a landmark in the treatment of erectile dysfunction. A considerable progress in understanding and treatment of erectile dysfunction in the 80s (2) paved the way to the introduction of sildenafil at the end of the 90s (3). This PDE5-inhibitor is a patient-friendly medication with a relatively high efficacy. The availability of it did increase the public interest in the subject of erectile dysfunction.

Warranty about the cardiovascular safety of sildenafil (4, 5) put the association between ED and cardiovascular disease (CVD) prominent on the agenda of physicians. It was already known that men with erectile dysfunction are more likely to have cardiovascular disease in their history (6). However, is erectile dysfunction also a sentinel to cardiovascular disease (6–9)? Answering this question is of clinical importance. A positive answer to this question implies men with erectile dysfunction are of special interest with respect to primary prevention of cardiovascular disease.

We aimed at quantifying the cardiovascular risk of men with erectile dysfunction in the population-based study presented here. We restricted the study to men seeking help for their complaints of erectile dysfunction. These men are relatively easily accessible for cardiovascular disease prevention, if required. We also took into account the introduction of sildenafil, as this might have changed both the population of men seeking help and the attitude of physicians towards men suffering from ED.

METHODS

All 16 million Dutch citizens are registered with a general practitioner. The Integrated Primary Care Information is a Dutch database with longitudinal medical records (10). The database contains identification information; notes; prescriptions; and indications for therapy, physical findings, referrals, hospitalisations, and laboratory values. Seventeen general practitioners delivered medical files to Integrated Primary Care Information at the start of 1992. This number increased to 150 general practitioners in 2001 (11). The 150 general practitioners provided records of about half a million patients.

Males born between 1924 and 1960 were selected for the present study. Those men for whom medical information was available for at least 1 year before entry into the study were enrolled. One year of information was required to disentangle incidence and prevalence of ED and its

From Radboud University Medical Centre, Nijmegen, the Netherlands (J.F., T.G.W.S., L.A.K., E.J.H.M.).
Address correspondence to: J. Frantzen, Free University Medical Centre, PO Box 7057, 1007 MB Amsterdam, the Netherlands. Tel.: +31 20 444 0272. E-mail: j.frantzen@vumc.nl.
This study was supported by an unrestricted grant from Pfizer International, New York.
Received December 3, 2004; accepted June 20, 2005.

360 Park Avenue South, New York, NY 10010
1047-2797/06/$–see front matter
doi:10.1016/j.annepidem.2005.06.047

Principles of Epidemiology for Public Health (EPID600)

Sources of error: Selection bias

Victor J. Schoenbach, PhD home page

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

www.unc.edu/epid600/

Excerpts from (allegedly) actual student history essays collected by teachers from 8th grade through college

Richard Lederer, St. Paul's School

http://www.edfac.usyd.edu.au/staff/souters/Humour/Student.Blooper.World.html

"Finally Magna Carta provided that no man should be hanged twice for the same offense."

"Another story was William Tell, who shot an arrow through an apple while standing on his son's head."

"It was an age of great inventions and discoveries. Gutenberg invented removable type and the Bible. Another important invention was the circulation of blood. Sir Walter Raleigh is a historical figure because he invented cigarettes and started smoking. And Sir Francis Drake circumcised the world with a 100 foot clipper."

Memória (in case you happen to speak Portuguese)

Dois amigos se visitam. Diz o anfitrião:
– Rapaz, descobri um remédio para memória fantástico.
– Ah, é? E como chama?
O anfitrião põe a mão na testa, pensa um pouco, vira-se para o visitante:
– Como é que chama mesmo aquela flor que tem espinhos?
– Rosa.
O anfitrião grita para dentro de casa:
– Ô Rosa, como é que chama mesmo aquele remédio pra memória, meu bem?

De Luciana V. Paiva, Osasco - SP, en Bom Humor Nosso E Dos Leitores", *Almanaque Brasil de Cultura Popular*. Maio 2001;3(26) (almanaquebrasil@uol.com.br). Exemplar de quem viaja TAM.

"To error is human"

- Science emphasizes systematic, repeatable, carefully-conducted observation
- Laboratory investigations are highly controlled, to minimize unwanted influences
- Human sciences must contend with many threats to validity

"To error is human"

Any epidemiologic study presents many, many opportunities for error in relation to:

- Selection of study participants
- Classification and measurement
- Comparison and interpretation

Accuracy, Bias, Error, Precision, Reliability, Validity

- *Accuracy*: lack of error – getting the correct result
- *Bias*: systematic error – independent of study size
- *Error*: discrepancy between the observed result and the true value
- *Precision*: absence of random error
- *Reliability*: repeatability of a measure
- *Validity*: absence of bias (or absence of all error)

Internal validity versus external validity

- *Internal validity*: whether the study provides an unbiased estimate of what it claims to estimate
- *External validity*: whether the results from the study can be generalized to some other population

"Random error"

- "that part of our experience that we cannot predict" (Rothman and Greenland, p. 78)
- Usually most easily conceptualized as sampling variability

Draw 1,000 random samples from population with 50% women

	Estimate of % Women	
Distribution	N = 100	N = 1,000
Highest	68	54.9
10%	above 56	above 51.9
25%	above 53	above 51.0
25%	below 47	below 48.9
10%	below 44	below 48.0
Lowest	33	45.0

Random error can be problematic, but . . .

- Influence can be reduced
  - – increase sample size
  - – change design
  - – improve instrumentation
- Probability of different types of influence can be quantified

However, many people do not understand

"96% of donors say they are 'very satisfied' or 'satisfied'"

"79% of donors have recommended the Charitable Gift Fund"

"Donors cite simplicity (70%), tax benefits (66%), and ease of use (62%) as the top three benefits"

Survey background: "Results were computed with 95% confidence level."

The Benefactor, fall 2001, p3 (Fidelity Investments Charitable Gift Fund newsletter)

Systematic error ("bias") is more problematic

- May be present without investigator being aware
- Sources may be difficult to identify
- Influence may be difficult to assess

Direction of bias - "which way is up?"

- *Positive bias* – observed value is higher than the true value
- *Negative bias* – observed value is lower than the true value
- *Bias towards the null* – observed value is closer to 1.0 than is the true value
- *Bias away from the null* – observed value is farther from 1.0 than is the true value

Classifying types of bias

- Selection bias – differential access to the study population
- Information bias – inaccuracy in measurement or classification
- Confounding bias – unfair comparison

Possible sources of selection bias

- Participant selection procedure, e.g., exposure affects case ascertainment ("detection bias") or control selection
- Differential unavailability due to death ("selective survival"), illness, migration, or refusal (nonresponse bias)
- Loss to follow-up / attrition / missing data

Examples

- Case-control study of physical fitness and heart attacks
- Cohort study of smoking cessation methods

Example: differential loss to follow-up (cohort study)

	Complete cohort		Observed cohort	
	Cases	Noncases	Cases	Noncases
Exposed	40	160	32	144
Unexposed	20	180	18	162
Total	60	340	50	306
Risk ratio	2.0		1.8	

Example: non-response bias in case-control study

	Target population		Study population	
	Cases	Noncases	Cases	Noncases
Exposed	200	20,000	180	200
Unexposed	400	40,000	300	400
Total	600	60,000	480	600
Odds ratio	1.0		1.2	

Selective survival can affect a cohort study

Natural history of HIV infection

1 Recruit participants at the time they become HIV infected

2 Recruit participants by means of a serologic survey to detect prevalent infections.

People who progress to AIDS and die relatively soon after infection will be underrepresented in study 2.

Example – cohort of survivors

Effect of hypertension in an elderly cohort

If people who die before becoming elderly were more vulnerable to end-organ damage from hypertension, then the cohort study may observe less morbidity associated with hypertension than would be observed if the study had enrolled younger participants.

Example – cohort of survivors

Effect of environmental tobacco smoke (ETS) in newborns

If ETS increases early fetal losses, possibly undetected, then the most susceptible fetuses may be unavailable to be part of the cohort study of effects of ETS on infants.

Conceptual framework

(Kleinbaum, Kupper, Morgenstern)

- Target population
- Actual population
- Study population
- "Selection probabilities"

Target population vs. actual population

- Target population – the population from which we <u>think</u> we are studying a sample
- Actual population – the population from which we are <u>actually</u> studying a sample
- Study population – a random sample from the actual population

Actual population vs. study population

- "Selection probabilities" – the probability that someone in the <u>target</u> population will be in the <u>actual</u> population

Selection probability (for each exposure-disease subgroup) = probability (target ⟶ actual)

Selection probabilities from the target population to the actual population

	Target population		Actual population	
	Exposed	Unexposed	Exposed	Unexposed
Cases	A	B	A^o	B^o
Noncases	C	D	C^o	D^o

alpha = A^o/A, beta = B^o/B, gamma = C^o/C, delta = D^o/D

Example of selection probabilities

Endometrial cancer and estrogen use

- Case-control studies found high OR's
- Horwitz and Feinstein claimed "detection bias": women may have asymptomatic tumors; if they have vaginal bleeding from estrogens their cancer will be discovered

Horwitz and Feinstein's argument

	Target population		Actual population	
	Estrogen	Unexposed	Estrogen	Unexposed
Endometrial cancer	2,000	8,000	1,800	900
Noncases	2,000,000	8,000,000	1,600,000	6,400,000
OR	1.0		8.0	

Horwitz and Feinstein's argument

	Actual population / Target population	
	Estrogen	**Unexposed**
Cases	alpha = 1,800 / 2,000	beta = 900 / 8,000
Noncases	gamma = 1,600,000 / 2,000,000	delta = 6,400,000 / 8,000,000

Horwitz and Feinstein's argument

Actual population / Target population

	Estrogen	Unexposed
Cases	alpha = **0.9**	beta = **0.11**
Noncases	gamma = 0.8	delta = 0.8

alpha/beta = 8 **>** gamma/delta = 1

Horwitz and Feinstein's solution

Actual population / Target population

	Estrogen	Unexposed
Cases	alpha = 0.9	beta = 0.11
Noncases	gamma = **0.36**	delta = **0.06**

alpha/beta = 8, gamma/delta = 6

Summary

- Epidemiologists differentiate between random error and systematic error ("bias").
- Bias: selection, information, and confounding.
- Bias can arise in any type of epidemiologic study.
- Ask: bias from what, how much, what effect?

Excerpts from (allegedly) actual student history essays collected by teachers from 8th grade through college
Richard Lederer, St. Paul's School

"The greatest writer of the Renaissance was William Shakespeare. He was born in the year 1564, on his birthday. He never made much money and is famous only because of his plays. He wrote tragedies, comedies, and hysterectomies, all in Islamic pentameter. Romeo and Juliet are an example of a heroic couplet. Romeo's last wish was to be laid by Juliet."

Excerpts from (allegedly) actual student history essays collected by teachers from 8th grade through college
Richard Lederer, St. Paul's School

"Writing at the same time as Shakespeare was Miguel Cervantes. He rote Donkey Hote. The next great author was John Milton. Milton wrote Paradise Lost. Then his wife died and he wrote Paradise Regained."