In-class exercise on Selection Bias — **Instructor Guide**

Background

The HIV epidemic in the United States began in the 1980's through three major sources of transmission: anal sexual intercourse, viral contamination of blood and blood products, and passage of virus through sharing and re-use of needles, syringes and other drug-use paraphernalia ("works") by injection drug users (IDU). As surveillance data and epidemiologic studies identified the key role of injection drug use in the spread of HIV, proposals to increase the availability of sterile injection supplies to reduce HIV transmission — a strategy characterized as "harm reduction" — emerged. Although Canada and a number of European countries experimented with "needle-exchange" programs (NEP) as a way of reducing the sharing of infectious needles and syringes, NEP has been a highly controversial issue in the United States, where they are prohibited by state laws and where the U.S. Congress has forbidden the use of federal funding for such programs.

Much of the controversy has centered around the concern that NEP send a "mixed message" about drug use and undermines the prohibitionist-stance of U.S. drug policy, thereby encouraging drug use. Controversy has also existed over whether NEP do in fact reduce HIV transmission. A fall 1993 report commissioned by the U.S. Centers for Disease Control concluded from studies available at that time that NEP were likely to reduce HIV transmission without increasing drug use rates, but the report was suppressed until the *Washington Post* obtained it in February 1995 (Lurie, 1997). A Congressionally-mandated report by the National Academy of Sciences reached a similar conclusion in September 1995, though three studies raising doubts about NEP effectiveness had not yet been published.

The first of these three studies appeared in the December 15, 1997 issue of the *American Journal of Epidemiology* (Bruneau J, Lamothe F, Franco F, Lachance N, Desy M, Soto J, Vincelette J, "High rates of HIV infection among injection drug users participating in needle exchange programs in Montreal: results of a cohort study", *Amer J Epidemiol* 146(12):994-1002), accompanied by an invited commentary by Peter Lurie ("Le mystere de Montreal") and a response from the authors. As the authors carefully note, their cohort study of active IDU in Montreal began a year prior to the introduction of NEP in Canada and was not designed for the purpose of evaluating NEP. Preliminary analyses of the data, however, indicated a possible increase in seroincidence of HIV in NEP users, so the authors carried out an extensive analysis of the relationship. This exercise is based largely on data from the Bruneau et al.

1. Random error: sample size, precision, and standard error

The study enrolled 1599 persons who had injected drugs during the preceding six months principally through self-referral or a detoxification facility. One hundred seventy one (10.7%) were HIV seropositive at baseline. How precise is the estimate of 10.7% HIV seropositivity at baseline? If Montrealers who have injected drugs during the six months before the study recruitment period are the population of interest and we assume simple random sampling, we can use introductory statistics to compute a confidence interval around this estimate of HIV seroprevalence. The width of this interval quantifies the variability inherent in selecting a sample of a given size. Conversely, the

reciprocal of the width of the confidence interval can be used as a measure of the precision of the estimate.

Compute the standard error and corresponding confidence intervals for the estimate of 10.7% if the sample size had been 900 or 2,500, and compute the corresponding confidence intervals. Describe the relationship between the sample size and relative width of the intervals. (The standard error is the square root of the variance of the estimate of the proportion [p(1-p)/n]. The formula for the 95% confidence interval an estimate of a proportion (p) in a large population from a simple random sample of size n is:

$$p \pm 1.96 \sqrt{p(1-p)/n}$$

Sample size n	$\sqrt{n}$	Prevalence p	p(1-p)	Standard error $\sqrt{[p(1-p)/n]}$	95% confidence limits Lower	Upper
900	30	0.107	0.095551	0.010304	0.087	0.127
1,600	40	0.107	(0.107)(0.893)	0.309113/40	0.092	0.122
2,500	50	0.107	0.095551	0.006182	0.095	0.119

Halving the standard error requires multiplying the sample size by <u>four</u>.

2. Nonresponse bias

Suppose that at the time the Bruneau et al. study began, the injection drug using population of Montreal, HIV seroprevalences for men and women in and not in treatment, and rates of participation in the study were as shown in the following table (e.g., 20% of the 3,620 male injection drug users participated in the study):

	HIV seroprevalence	Population size	Participation rate	Participants	HIV seroprevalence
Total	**0.0834**			1,599*	0.107*
Males		4406		1274*	0.144
Females		2746		325*	0.084
In treatment					
Males	0.150	786	0.70	550	0.150
Females	0.090	72	0.80	58	0.090
Not in treatment					
Males	0.100	3620	0.20	724	0.100
Females	0.041	2674	0.10	267	0.041

* All numbers in the table are hypothetical except these.

Under this scenario, what would the crude seroprevalence be in the IDU population of Montreal? Explain why the seroprevalence estimate from the Bruneau study differs from this hypothetical population seroprevalence. (If you need a hint, check the end of the exercise.)

3. Loss to follow-up

There were 974 participants who were seronegative at baseline and who were followed up for a median of 15.4 months. 89 seroconversions were noted, for an overall incidence of 5.1 per 100 person-years. (Note: the 15.4 months is a median, not a mean, so 89/(974*15.4/12) does not equal 5.1 per 100. The appropriate analysis of these data is based on person-time, but for simplicity the following questions ask for incidence proportions.)

a. What was the cumulative incidence of seroconversion?

Answer: CI = 89/974 = 9.1%

377 initially-seronegative participants were lost to follow-up. These 377 baseline participants differed from the 974 participants with follow-up data in several characteristics, including gender (81% vs. 74% male), income (11.5% vs. 21% reported lower income), and getting syringes and needles from a drug dealer (57% vs. 33%). However, the participants lost to follow-up were no more or less likely to have attended a NEP. (Note: the remaining 77 baseline participants were excluded since they had been recruited too late in the study to have at least three months of follow-up.)

Suppose that the cumulative HIV sero-incidence in the 377 participants lost to follow-up was twice that in the 974 participants with follow-up, and that this ratio held for both NEP users and nonusers.

b. Under these assumptions, what was the actual cumulative incidence of seroconversion among the 377+974 initially-seronegative participants?

Answer: Using the unrounded cumulative seroincidence of 89/974,

[89+377 x 2 x (89/974)]/(974+377), the actual CI was:

[(974)(9.1%) + (377)(2)(9.1%)] / (974+377) = 11.7%

c. Under these assumptions, was the seroconversion ratio for NEP to non-NEP persons biased by the loss to follow-up? Explain (if you prefer arithmetic to algebra, try using the hypothetical numbers in the following table):

	Subjects with follow-up data			Subjects lost to follow-up		
	Baseline	Incident cases	CI	Baseline	Incident cases	CI
Users	322	50	0.15	125	39	0.30
Nonusers	652	39	0.06	252	30	0.12
Total	974	89	0.09	377	69	0.18

If loss to follow-up did not differ by NEP, then the proportion of NEP and non-NEP who were lost to follow-up are both equal at 377/(974+377) = 28%. So the true cumulative incidence would be:

$0.72\ CI_{NEP} + 0.28*2\ CI_{NEP} = 1.28\ CI_{NEP}$ in NEP attenders and

0.72 $CI_{non\text{-}NEP}$ + 028*2 $CI_{non\text{-}NEP}$ = 1.28 $CI_{non\text{-}NEP}$ in non-NEP.

Therefore the true CIR = 1.28 CI_{NEP} /1.28 $CI_{non\text{-}NEPs}$ = CI_{NEP}/$CI_{non\text{-}NEP}$, the same as the observed seroincidence.

With numbers, (50+39)/(322+125) = 0.198 in non-NEP;
(39+30)/(652+252) = 0.0768 in NEP attenders,
CIR = 0.198/0.0768 = 2.6 = 0.15/0.06

Note, however, that there would be bias in the odds ratio and the incidence rate ratio.

4. Berkson's bias

Switching gears, consider a case-control study to see whether diabetes is a risk factor for pneumonia. Cases are persons hospitalized for viral pneumonia (bacterial, viral, or mycoplasma). Controls (N=220) case are selected by random-digit dialing. Suppose that the number of persons with and without diabetes in the population and the number of new cases of pneumonia that develop during one year are:

	Entire population	Incidence rate per 10,000 py	Rate of hospitalization	Hospitalized for pneumonia
Persons with diabetes	10,000			
Develop pneumonia	200	200	0.40 or 0.80	**160**
Persons without diabetes	100,000			
Develop pneumonia	1,000	100	0.40	400

a. What is the incidence rate ratio (i.e., the IDR) for pneumonia comparing persons with and without diabetes?

IDR = (200/10,000) / (100/100,000) = 20 / 10 = 2.0

b. What odds ratio would be obtained by the above-described case-control study if 40% of all pneumonia cases are hospitalized (in computing the odds of exposure in the control group, use the entire population figures).

OR = (80 × 200) / (400 × 20) = 2.0

c. Suppose that primary care physicians are twice as likely to hospitalize a pneumonia patient who also has diabetes, so that 80% of pneumonia patients with diabetes are hospitalized. What odds ratio would the case-control study estimate in this scenario?

OR = (160 × 200) / (400 × 20) = 4.0

d. Would the odds ratios in b. and c. be greater or lower if controls had instead been selected from hospitalized persons admitted for reasons other than infection (e.g., cardiovascular, genito-urinary, gynecological, and trauma patients)?

Since diabetes is associated with various disorders that may lead to hospitalization, the prevalence of diabetics among hospitalized patients will be greater. Therefore the estimated OR's will be lower.

Hint for question 2: The crude HIV seroprevalence is a weighted average of the seroprevalence for each subgroup, weighted by the proportionate subgroup sizes.