BIAS IN STUDY DESIGN

Two types of study bias

1. Systematic Error–Bias

2. Random Error–random variability in data not caused by systematic error. Random error decreases as sample size increases.

Systematic Error/Bias

–cannot be reduced by increasing sample size

Many types of bias have been discussed (see Sackett) but can be reduced to two broad types of bias:

Selection bias

Information bias

SELECTION BIAS

Many types of selection bias–

Concern is that subjects have different probability of being selected according to *exposures* or *outcomes* of interest, creating a biased measure of association (i.e. odds ratio, relative risk)

TWO MAIN TYPES OF SELECTION BIAS

Medical Surveillance Bias

Berkson Bias

Medical surveillance bias

More likely in case control studies where cases are ascertained through medical clinics, hospitals. If clinical visits are associated with the exposure, sub-clinical cases are more likely to be detected among those with the exposure than those without the exposure.

Ex. (From Szklo): Case Control Study of Oral Contraceptive Use and Diabetes–OC users more likely to have medical visits, resulting in higher probability of sub-clinical disease being detected. Any association with OC use and diabetes would be an overestimate of risk because sub-clinical diabetics with no OC use would have a lower probability of being selected.

Also called "unmasking" or (detection signal) bias in Sackett

Exposure "causes a sign or symptom which precipitates a search for the disease"

Ex. Post menopausal estrogen users and endometrial cancer– Stage 1 endometrial cancers were more likely to be detected among patients who presented with bleeding (caused by estrogens)–so cases were over represented among estrogen users. When controls were selected from patients who had hysterectomies/dilatation and curettage, OR was reduced because more stage 1 tumors were found among those originally presenting without disease.

This bias can be reduced by including only late-stage tumors, which are likely to be diagnosed equally in estrogen/non-estrogen users.

Possible solutions to medical surveillance bias:

1) Only use cases and controls who have undergone similar detection procedures–

However, some exposures may cause similar diseases in same organ. Ex. Study of lung cancer, persons with chronic pulmonary disease would comprise a large portion of controls (detection procedures for lung cancer would also detect CPD). CPD related to smoking, so smoking would be over-represented in control series, and OR for smoking and lung cancer would be muted.

2) Concurrent Prospective Study–ascertain cases as they occur through regular medical surveillance regardless of exposure status (mask exposure status when ascertaining the outcome)

3) Obtain information on type of medical care received to determine whether frequency of medical care may affect diagnosis of disease, can stratify by this variable

4) Stratify by disease severity in the analysis (if exposed cases more likely to be diagnosed in early stages)

Selection Bias (continued)

Berkson bias

Cases and/or controls selected from hospitals.

If hospital based cases/controls have different exposures than population based cases/controls, OR will be biased (could be over or underestimated)

Ex. If hospital based controls are less likely to have exposures of interest than population they are supposed to represent, OR will be over-estimated.

Ex. Case control study of pancreatic cancer and coffee drinking. Controls were selected from gastroenterologist's patients in same hospital. However, GI patients are less likely to drink coffee than the rest of the population because of their disease. Hence the OR for coffee drinking was artificially increased due to the under-representation of coffee drinkers among controls.

Solution–use population based controls, or controls with diseases not related to the exposure.

Berkson Bias (continued)

Sackett example: Hospitalized cases more likely to have factor than non-hospitalized cases, OR overestimated.

(If hospitalized cases represent all or most of population cases, than Berkson bias only applies to hospital based controls—i.e most cancers dx in hospitals)

SELECTION BIAS IN CASE CONTROL STUDIES

Bias only occurs when sampling fraction for cases and controls is related to *exposure*

Example of No Information Bias Occurring
Table 4-1 Szklo

Total Population

Risk Factor Present	Cases	Controls
Yes	**500**	**1800**
No	**500**	**7200**

Exposure Odds Cases 500/500= 1:1
Exposure Odds Controls 1800/7200= 1:4
OR=

$$\frac{\frac{500}{500}}{\frac{1800}{7200}} = 4.0$$

Sample of Total Population

Risk Factor Present	50% of Cases	10% of Controls
Yes	**250**	**180**
No	**250**	**720**

Exposure odds cases : 250/250= 1:1
Exposure odds controls: 180/270=1:4
OR=

$$\frac{\frac{250}{250}}{\frac{180}{720}} = 4.0$$

Example of Sampling Bias Associated with Exposure

Risk factor	Cases	Controls
Present	.60 x 500=300	.10 x 1800=180
Absent	.40 x 500=200	.10 x 7200=720

Higher proportion of exposed cases are selected than non-exposed cases, resulting in a biased odds ratio

Exposure odds in cases: 300/200= 1.5:1.0
Exposure odds in controls: 180/720=1:4

Odds Ratios :

$$\frac{\frac{300}{200}}{\frac{180}{720}} = 6.0$$

SELECTION BIAS (continued)

Example of Same Level of Sampling Bias in Selecting Cases and Controls

Risk factor	Cases	Controls
Present	.60 x 500=300	.136 x 1800=245
Absent	.40 x 500=200	.091 x 7200=655
TOTAL	500	900

Exposure Odds Cases : 300/200= 1.5: 1.0

Exposure Odds Controls: 245/655=1.0:2.67

Exposure Odds Ratio:

$$\frac{\frac{300}{200}}{\frac{245}{655}} = 4$$

Bias is similar in cases, controls so OR is not biased:

.60/.40=1.5
.136/.091=1.5

Sampling fraction in exposed cases, controls, is 1.5 times sampling fraction in unexposed cases, controls. “Compensating bias”

SAMPLING BIAS (continued)

SAMPLING BIAS IN CASES

If all cases are selected, no sampling bias in cases

Otherwise, medical surveillance bias, Berkson bias possible sources of bias in cases

SAMPLING BIAS IN CONTROLS

Berkson Bias--Hospital controls–Usually higher response rates than population controls, but more likely to have sampling bias (hospital controls likely to have different exposures than population controls)

Population controls–Less sampling bias, but usually higher refusal rates

Friend controls: Overmatching possible–friends may have similar exposures Ex. Smokers may have friends who are smokers--

TO REDUCE SAMPLING BIAS

Cases/controls selected from same source (i.e. same clinic, HMO, screening clinic etc)

Szklo—Screening programs for breast cancer; cases, controls selected from these screening programs should have similar risk factors (women who go to screening clinics more likely to have higher prevalence rates for certain risk factors, such as family history of breast cancer)

However, cannot assume that sampling bias will be similar in cases, controls. Referral patterns may be different for different diseases

Best to sample cases, controls from same defined population, but not hospital or clinic based

i.e. Geographic based (all Cook County residents), employees of same company, etc

Example:

Case control study of pancreatic cancer–

Define cases as all Cook County residents, diagnosed between time x and time y, with pathologic confirmation of cancer, between age 35 to 74

Obtain cases through hospitals/clinics, state cancer registry

Obtain population controls through Random Digit Dialing, HCFA

INFORMATION BIAS

Collection of erroneous study data due to:

Invalid /imprecise study measures:

Validity: Does your test/variable measure what it is supposed to measure–can be disease or exposure

Sensitivity–ability to identify those who have the disease/factor

Specificity-ability to identify those without the disease/factor

Reliability: Are the results replicated with repeated testing? Can be biologic sample, a questionnaire regarding exposures (diet, smoking, workplace etc)

Errors in data collection procedures–can take many forms!

EXPOSURE IDENTIFICATION BIAS

Are exposures misclassified?

Case control studies have more potential for misclassification of exposure because exposure status is collected *retrospectively*.

Concurrent Cohort Studies can collect exposure data prospectively so there is less potential for error. (Although some exposure may be collected retrospectively, i.e. information on exposures before the study started, such as prior occupational, dietary, environmental exposures etc)

Most common type of information bias is *RECALL BIAS*

RECALL BIAS

Recall bias occurs when cases and controls recall exposures differently.

For example, cancer cases may try harder to recall prior exposures because they think they might be related to their disease. Parents of children with birth defects may try harder to recall any drugs, exposures they had during pregnancy than parents of children without birth defects. Odds ratios in these examples would be artificially inflated.

WAYS TO AVOID RECALL BIAS

1) Verification of exposure data

Not always possible.

Ex. Medical record reviews for prior prescription drugs, procedures, diagnoses, review of work records for exposure data, etc. Depends on the availability and accuracy of these records. Also verification of prescriptions does not tell you whether the person actually took the drugs, working in a certain department may not tell you what specific exposures were in that department etc.

2) Use controls with other diseases

"Rumination bias," the idea that people with diseases will think harder about their prior exposures than disease free people, might be reduced if controls are people with diseases other than the one under study. However, those with serious diseases like cancer, may have a different rumination process than those diagnosed with other, less serious diseases. In addition, those with other diseases are more likely to have different exposures than the general population (sampling bias).

So you may reduce one type of bias (recall) but introduce another type (sampling)! Some studies of cancer have used other types of cancer as controls (especially using cancer registries), but the cancer controls should exclude cancers that are related to the exposure of interest. For example, a study of smoking and lung cancer would not include control cancers also related to smoking, such as bladder, pancreas.

WAYS TO REDUCE RECALL BIAS (continued)

3) Use "OBJECTIVE" Markers of Exposure

Biologic, genetic markers, *if available*, may be used to verify exposure.

Ex. Szklo–DNA repair abilities as a genetic marker for susceptibility to ultra violet light-induced melanoma skin cancer.

***Environmental markers*– most represent current, not past exposure–cotinine measures recent smoking, chemical exposures with short half lives cannot be measured (i.e, trichloroethylene exposure). PCBs have long half lives and can be measured in blood but different types of PCBs last longer (congeners) so you may only be measuring less harmful PCBs.**

4) CHANGE STUDY DESIGN

Use nested case control study–main exposures are collected at baseline of cohort study (including stored serum), which can be analyzed at a later point in study. Ex. Nurses health study, used nested case control design, analyzed PCBs from stored serum collected at baseline. Dietary data can be collected at baseline, and throughout study.

OTHER TYPES OF INFORMATION BIAS

Interviewer bias:

Occurs when interviewers conduct interviews differently for cases, controls. Can occur when interviewers know the case status of the subject. For example, an interviewer may probe cases more than controls (are not usually aware that this occurs)

Ways to avoid:

Blind interviewers to subject status (case control)–not always possible–cancer cases may be sick, impossible to hide their disease status.

Train staff, standardize data collection procedures, quality control etc. (Standard survey research procedures)

RESPONSE BIAS

–Occurs when response rates vary according to exposure status

A biased Odds Ratio can result from *differential response rates.* If a higher proportion of exposed cases then exposed controls participate in your study, bias will also occur. In a cohort study, similar bias can occur due to differential losses to follow-up. If disease/exposure is related to drop outs, the RR will be biased.

Ex. A cohort study examining smoking and cancer outcomes. Smokers are more likely to have health problems than non-smokers, and may be more likely to drop out due to illness (i.e. emphysema). These smokers are also more likely to develop the disease of interest (i.e. lung cancer) but these events will not be counted. Thus, smokers who develop lung cancer will be under-represented in your study and the RR will be an underestimate.

MISCLASSIFICATION OF EXPOSURE –HOW DOES IT AFFECT THE ODDS RATIO?

DIFFERENTIAL VS NON DIFFERENTIAL MISCLASSIFICATION OF EXPOSURE/DISEASE

NON DIFFERENTIAL EXPOSURE MISCLASSIFICATION:

If the exposure misclassification is NOT associated with the disease

(Misclassification of exposure is similar in cases and controls)

Result–when there is a dichotomous exposure, the result will be biased towards the null (no association)

Non- differential exposure classification example

From Rothman Table 5-2

Non differential classification

	Correct Classification		20% of no say yes		20% of no say Yes 20% of yes Say no	
	High fat diet		High fat diet		High fat diet	
	No	Yes	No	Yes	No	Yes
Heart attack cases	450	250	360	340	410	290
Controls	900	100	720	280	740	260
Odds ratio	5.0		2.4		2.0	

Similar misclassification in cases and controls leads to underestimate of odds ratio.

Non differential misclassification of a *dichotomous* exposure will always bias the odds ratio towards the null. For non-dichotomous exposures, the bias could go in either direction.

Degree of misclassification also increases with decreasing exposure prevalence in controls, and when sample size is much greater in controls than cases. (See Table 4-5 in Szklo p.146)

EFFECTS OF *DIFFERENTIAL* EXPOSURE MISCLASSIFICATION

If the exposure misclassification is associated with the disease, it can either *exaggerate* or *underestimate* an effect.

Szklo Table 4-8 (Nurses Health Study data)

	Pre melanoma Diagnosis “Gold standard”		**Post melanoma Diagnosis**	
Tanning ability	**Cases**	**Controls**	**Cases**	**Controls**
No tan to light tan	**9**	**79**	**15**	**77**
Medium to dark tan	**25**	**155**	**19**	**157**
Odds ratio	**0.7**		**1.6**	

Cases inaccurately recall tanning ability compared to their initial assessment pre diagnosis, leading to an over estimate of the odds ratio.

(Of 15 cases in “exposed” category post melanoma diagnosis, 8 were true positives, 7 were false positives)

Example of Differential Misclassification of *Disease* Status in a Cohort Study

Differential follow up in a cohort study

Ex. Cohort study examining effect of smoking on emphysema

Diseases are ascertained by asking subjects for self reports of medical diagnoses, not by any medical exams conducted in the entire cohort.

Emphysema may be more likely to be diagnosed in smokers than non smokers, because physicians will be more likely to look for it in non smokers, and smokers are likely to have more contact with physicians due to other health problems related to smoking.

Relative Risk will be overestimated, because disease will be missed more in non-smokers.

Solution–conduct medical exams in all subjects—however, very expensive in a large cohort study

OTHER TYPES OF BIAS

Incidence-Prevalence Bias in Cross sectional studies

Prevalence –function of incidence, duration

If exposed prevalent cases have a different duration than non-exposed prevalent cases, the Prevalence Rate Ratio (PRR) will be biased.

Ex. Cases with severe emphysema more likely to smoke, have higher fatality than cases with less severe emphysema, so the prevalence of emphysema in smokers will be underestimated compared to incidence

Point Prevalence Odds= $\frac{\Pr ev}{1-\Pr ev} = Inc \times Dur$

Or, Prevalence=Inc x Dur x (1-Prev)

Prevalence Rate Ratio (PRR)=

$$\frac{Inc_{\exp}}{Inc_{non-\exp}} \times \frac{Dur_{\exp}}{Dur_{non-\exp}} \times \frac{1-\Pr ev_{\exp}}{1-\Pr ev_{non-\exp}}$$

If Duration differs between exposed, non-exposed, dur/dur NE 1, PRR will be biased

Solution–use incident cases!

Temporal Bias

Primarily occurs in Cross Sectional Studies

Occurs when you don’t know the temporal sequence, and outcome and exposure data are collected at the same time

Ex. High blood pressure, elevated serum creatinine (marker of kidney failure)

Don’t know if elevated creatinine came before or after elevated blood pressure

Temporal Bias in Case Control Studies

Can occur when exposure data measured after dx–i.e. biologic samples collected in a case control study

Some biologic exposure measurements only measure recent exposure, i.e. TCE exposure

Szklo ex. HBV and idiopathic aplastic anemia (IAA):

Serum samples for HBV collected after onset of disease–Cases with undiagnosed IAA may have received blood transfusions of blood contaminated with HBV before dx of IAA.

Descriptive studies/Ecologic studies

Autism rates in Brick Township NJ. High incidence of autism in this town may be due to parents moving there because of special school for autistic children.

Screening Bias

Biases that may occur in studies of evaluation of screening interventions

Selection Bias

Incidence- Prevalence Bias

Length bias

Lead time bias

SCREENING BIAS (continued)

Selection bias

occurs when subjects are not randomized into screened/non-screened groups

Persons who attend screening programs usually of higher SES than those who don’t attend screening programs. Outcomes may be due to SES, not the screening program. Solution–randomize groups.

Incidence Prevalence Bias

Prevalent cases detected at first screening more likely to be long term survivors with better prognosis. Solution–use cases detected at subsequent screens.

SCREENING BIAS (continued)

Length bias

Occurs when cases detected by screening program have better prognosis than cases detected between screening programs

Screening detected cases may have a longer detectable pre-clinical phase (DPCP) then those detected between screenings, hence a better prognosis that is not related to the screening

Ex. Interval cases of breast cancer, (detected between screenings) had more rapidly growing tumors, hence a higher case fatality rate

Analysis of screening program needs to take into account any differences in the DPCP

Lead time bias

Time in which diagnosis is advanced by screening

Time between early dx and when dx would have been made without screening. Bias occurs when estimating survival from the time of diagnosis instead of the time when the disease would normally have been detected.

(If you know stage of disease, you can calculate survival differently according to stage)