2012 ISPE Mid-Year Meeting
Introduction to Pharmacoepidemiology
April 22, 2012

Case-Control Studies

Tobias Gerhard, PhD
Assistant Professor, Ernest Mario School of Pharmacy
Institute for Health, Health Care Policy, and Aging Research
Rutgers University

RUTGERS
UNIVERSITY

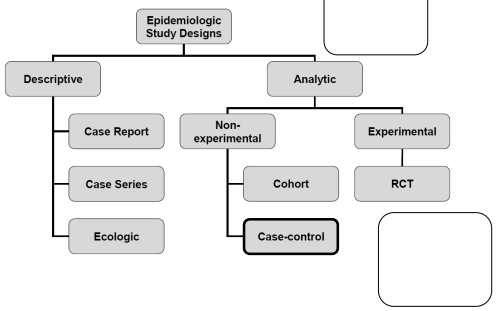
Conflict of Interest

- The views and opinions expressed in this presentation are solely mine and do not represent the position or opinion of ISPE or any other institution.
- I have no conflicts of interest to declare.

Outline

- Overview and general principles
- Measures of association
- Variants of the case-control design
  - Incidence density case-control studies
  - Case-cohort studies
  - Cumulative case-control studies
- Final thoughts and take-home points

Epidemiologic Study Designs



Preface

- Case-control studies are difficult to understand
- Many misconceptions prevail
- If this is new to you, don't be dissappointed if you don't follow-everything the first time

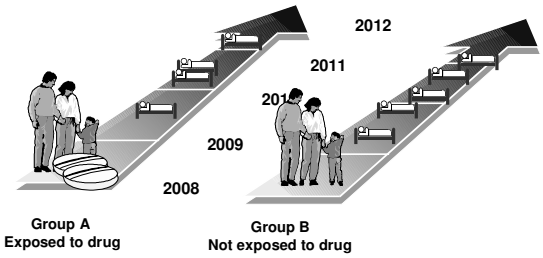
Case-Control Studies

Definition

The observational epidemiologic study of persons with the disease of interest and a suitable control group of persons without the disease. The relationship of an attribute to the disease is examined by comparing the diseased and non-diseased with regard to how frequently the atribute is present.

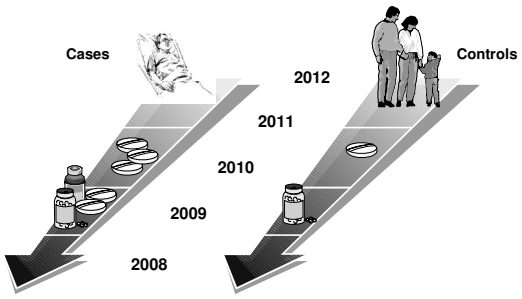
John M. Last, Dictionary of Epidemiology

Cohort Studies
Design



DISEASE FREQUENCY IN EXPOSED VS. UNEXPOSED

Case-Control Studies
Design



FREQUENCY OF DRUG EXPOSURE IN CASES VS. CONTROLS

Case-Control Studies
Conduct

1. Define the *source population* for the study (hypothetical study population in which a cohort might have been conducted). Also referred to as the *study base*.
2. Define exposure(s) and outcome of interest.
3. Identify *cases* (subject with disease of interest) and determine their exposure status.
4. Instead of determining the exposure status for the remainder of the source population (as would be the case in a cohort study), exposure status is determined only for a much smaller number of patients sampled from the source population. These *controls* provide an *estimate of the prevalence of exposure (and covariates) in the source population.*
5. Calculate and interpret the exposure odds ratio.

Why conduct a case-control study?

- A case-control study can be conceptualized as a **more efficient version of a corresponding cohort study**.
  - Disease/outcome of interest is rare
  - Long latency or induction period
  - Exposure or confounder data expensive to collect
- Allows study of multiple exposures

Efficiency

Thiazides and femur fractures? A tale of two studies.

	Cohort Study (Feskanich et al., Osteoporos Int 1997)	**Case-Control Study** (Herings RM, et al. J Clin Epidemiol 1996)
Source Population	**83,728** women (36-61 years) followed over 10 years with biennial questionnaires	**300,000** Dutch residents included in the PHARMO database
Femur Fractures	**251**	**386**
Exposure ascertainment of underlying cohort	**83,728**	**386** randomly selected controls (matched for age, sex, pharmacy and GP)
Effect Estimate	RR: 0.69 (0.48-0.99)	OR: 0.5 (0.3-0.9)

Disadvantages of Case-Control Studies

- Cannot calculate incidence of disease in the population (unless information of the underlying cohort is accessible)
- Generally inefficient for rare exposures
- Generally limited to one outcome of interest
- **Greater *potential* for bias:**
  - sampling of controls not independent of exposure status
  - exposure or confounder ascertainment influenced by occurrence of the study outcome

Defining the Source Population

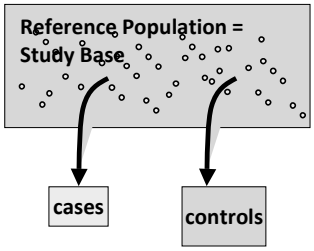
- **Ideal:** *population-based*; the source population can be precisely described and identified (primary study base)
  - Automated databases
  - Population registries of defined geographic regions
    → **Nested case-control study**
- However, it is often not possible to identify the source population explicitly (*secondary study base*). In this instance, the source population is defined from (secondary to) a given set of cases.

Case Selection

- **Ideal:** *all incident cases in the source population*
  - Sampling of cases is possible (as long as it is a representative sample, independent of exposure)
- Convenience samples of case-series (e.g., all incident cases in a specific hospital) are possible (and fairly common) but pose problems for the precise identification of the source population (secondary study base)
- Case-control studies of prevalent cases require strong assumptions and are generally problematic

Control selection

Controls should be chosen from the **same base population** as the cases (the population that gave rise to the cases)



Representative sample of reference population from which cases originated

Control Selection

- **Purpose:** to provide an unbiased estimate of the prevalence of exposure (and covariates) in the source population
- **Ideal:** (1) d*irect random sampling from the source population (independent of exposure)*
- If cases are recruited from a convinience sample and the source population is not identifiable, random sampling is not possible. Approaches to simulate random sampling include:
  - Hospital-based controls, neighborhood controls, random digit dialing

Most difficult element in case-control studies (particularly those with secondary study bases)

Example of a Secondary Base Case-Control Study

- Cases are identified in one hospital
- Source population (study base): All individuals who, if they had developed the outcome of interest in the same time period as the cases, would have been admitted to the same hospital.
  → **not precisely identifiable**
- Conceptualize patients admitted to the same hospital for other conditions (not thought to be related to the exposure of interest) as a representative sample of the hypothetical source population.

Great potential for selection bias

Ascertainment of Exposure and Confounding Variables

- Data collection
  - interview, questionnaire, etc
  - medical records, birth certificates, etc
  - automated datasources (claims, EMRs, etc)
- **Must be carried out identically for cases and controls**
  - Ask same questions in the same manner
  - When possible, researchers should be blind to the case/control status of the interviewer
  - Generally less problematic when historic or automated records are used

Measures of Association

- **The measure of association in case-control studies is the exposure odds ratio (OR)**
- Calculation of the odds ratio is the same for all types of case-control studies
- Depending on the sampling paradigm used to select the control subjects (and in some cases additional assumptions), the exposure odds ratio from a case-control study estimates different effect measures in the underlying cohort. *More on this later...*

Odds and Risk

- Assume the following 2x2 table from a cohort study

	Disease	Disease-Free	Total	Person-Time
Exposed (1)	A_1	B_1	N_1	T_1
Unexposed (0)	A_0	B_0	N_0	T_0
	A	B	N	T

- **Risk:** Probability that the event of interest (A) occurs:
  Risk of disease = A/N = A/(A+B)
- **Odds:** Ratio between the probability that the event of interest occurs to the probability that it does not:
  Odds of disease = A/(A+B) / B/(A+B) = **A/B**
  Odds of exposure = $N_1/(N_1+N_0)$ / $N_0/(N_1+N_0)$ = N_1/N_0

Odds Ratio

	Disease	Disease-Free	Total	Person-Time
Exposed (1)	A_1	B_1	N_1	T_1
Unexposed (0)	A_0	B_0	N_0	T_0
	A	B	N	T

- **Odds ratio** = ratio between two odds (*cross-product ratio*)
- **Exposure OR** = $(A_1/A_0)/(B_1/B_0) = (A_1 \cdot B_0)/(A_0 \cdot B_1)$
- **Incidence OR** = $(A_1/B_1)/(A_0/B_0) = (A_1 \cdot B_0)/(A_0 \cdot B_1)$
- **Exposure OR= Incidence OR**

Interpretation of the Odds Ratio

- **OR = 1**: no association between outcome and exposure (same odds of exposure in cases and controls = same odds of disease in exposed vs. unexposed)
- **OR >1**: exposure is associated with increased risk for outcome (greater odds of exposure in cases than controls = greater odds of disease in exposed vs. unexposed)
  Harmful Effect
- **OR <1**: exposure is associated with reduced risk for outcome (lower odds of exposure in cases than controls = lower odds of disease in exposed vs. unexposed)
  Protective Effect
- **Always consider the confidence interval!**

Interpretation of the Odds Ratio

Rare disease assumption: Only in case of a rare disease is the exposure ***odds ratio*** calculated from a cohort a valid approximation of the ***risk ratio*** for the same exposure.

	D +	D -
E +	200	9800
E -	100	9900

$RR = \frac{200/(200+9800)}{100/(100+9900)} = 2.00$

$OR = \frac{200 \cdot 9900}{100 \cdot 9800} = 2.02$

	D +	D -
E +	5000	5000
E -	2500	7500

$RR = \frac{5000/(5000+5000)}{2500/(500+5000)} = 2.00$

$OR = \frac{5000 \cdot 7500}{2500 \cdot 5000} = 3.00$

Rare Disease Assumption (formal)

Relationship of the relative risk to the odds ratio

$$RR = \frac{I_e}{I_u} = \frac{\left(\frac{a}{a+b}\right)}{\left(\frac{c}{c+d}\right)} = \frac{\left(\frac{a}{a_{\to 0}+b}\right)}{\left(\frac{c}{c_{\to 0}+d}\right)} = \frac{\left(\frac{a}{b}\right)}{\left(\frac{c}{d}\right)} = \frac{ad}{bc} = OR$$

I_e = Rate of new events among exposed
I_u = Rate of new events among unexposed
Under the ***rare disease assumption***, events in the denominator are negligible (=0), because they have no important influence on the size of the denominator.

	D +	D -
E +	a	b
E -	c	d

Control Selection Paradigms

The controls provide an estimate of the prevalence of exposure and covariates in the source population. Controls can be selected from members of the source population who were

- noncases at the time that each case occurs (i.e., in proportion to the person time accumulated by the cohort) → **incidence density case-control study**
- noncases at the beginning of the study's follow-up period → **case-cohort study**
- noncases at the end of the study's follow-up period → **cumulative case-control study** ("traditional" case-control study)

Control Selection Paradigms

Source cohort

	Disease	Disease-Free	Total	Person-Time
Exposed (1)	A_1	B_1	N_1	T_1
Unexposed (0)	A_0	B_0	N_0	T_0
	A	B	N	T

Control Selection Paradigms

Source cohort

	Disease	Disease-Free	Total	Person-Time
Exposed (1)	A_1	B_1	N_1	T_1
Unexposed (0)	A_0	B_0	N_0	T_0
	A	B	N	T

Sampling for case-control study

Case-Control Study	**Cases**	**Controls**
	A or a*	b

*cases may or may not be a sample of the cases in the source cohort

Control Selection Paradigms

Source cohort

	Disease	Disease-Free	Total	Person-Time
Exposed (1)	A_1	B_1	N_1	T_1
Unexposed (0)	A_0	B_0	N_0	T_0
	A	B	N	T

Sampling from the population at risk (noncases) at the time that each case occurs

Case-Control Study	**Cases**	**Controls**
	A	b

Control Selection Paradigms

Source cohort

	Disease	Disease-Free	Total	Person-Time
Exposed (1)	A_1	B_1	N_1	T_1
Unexposed (0)	A_0	B_0	N_0	T_0
	A	B	N	T

Sampling from the population at risk at the beginning of the study period

Case-Control Study	**Cases**	**Controls**
	A	b

Control Selection Paradigms

Source cohort

	Disease	Disease-Free	Total	Person-Time
Exposed (1)	A_1	B_1	N_1	T_1
Unexposed (0)	A_0	B_0	N_0	T_0
	A	B	N	T

Sampling from the population at risk (noncases) at the end of the study's follow-up period

Case-Control Study	**Cases**	**Controls**
	A	b

Control Selection Paradigms

Sampling Design	Controls estimate prevalence of exposure proportional to	Effect measure estimated by case-control odds ratio
Incidence density case-control study	*Person time* in the source population	**Relative risk (rate ratio)** in underlying cohort
Case-cohort study	*Frequencies* of exposed and unexposed in the source population at start of study	**Relative risk (risk ratio)** in underlying cohort
Cumulative case-control study	*Frequencies* of exposed and unexposed noncases in the source population at end of study	**Incidence odds ratio** in underlying cohort (*with additional assumptions, this approximates the RR*)

Odds Ratio in Incidence Density Case-Control Studies

Recap:

- Incidence exposed = # exposed cases / exposed person time

  $$\mathbf{I_1 = A_1/T_1}$$

- Incidence unexposed = # unexposed cases / unexposed person time

  $$\mathbf{I_0 = A_0/T_0}$$

- Rate ratio (relative risk) = incidence exposed / incidence unexposed

  $$\mathbf{RR = (A_1/T_1) / (A_0/T_0)}$$

In case-control studies, neither exposed nor unexposed person time (T_1, T_0) are directly measured.

	Disease	Disease-Free	Person-Time
Exposed (1)	A_1	B_1	T_1
Unexposed (0)	A_0	B_0	T_0

Odds Ratio in Incidence Density Case-Control Studies

- Rate ratio (relative risk) = incidence exposed / incidence unexposed

  $$\mathbf{RR = (A_1/T_1) / (A_0/T_0)}$$

  can be rearranged to

  $$\mathbf{RR = (A_1/A_0) / (T_1/T_0)}$$

= odds of exposure in diseased / ratio of exposed to unexposed person-time

In case-control studies, neither exposed nor unexposed person-time (T_1, T_0) is directly measured... but

If controls (b) are selected to reflect the exposure distribution of person-time in the source population →

$$\mathbf{b_1/b_0 = T_1/T_0}$$

	Disease	Disease-Free	Person-Time
Exposed (1)	A_1	B_1	T_1
Unexposed (0)	A_0	B_0	T_0

Odds Ratio in Incidence Density Case-Control Studies

If **RR = $(A_1/A_0) / (T_1/T_0)$**
and **$b_1/b_0 = T_1/T_0$**
we can replace **T_1/T_0** with **b_1/b_0**
→ **RR = $(A_1/A_0) / (b_1/b_0)$ ... the case-control exposure odds ratio**

If controls are selected in proportion to the person-time they contribute to the underlying cohort (and independently of exposure status), the exposure odds ratio from the case-control study directly estimates the relative risk (incidence rate ratio) in the source population.

	Disease	Disease-Free	Person-Time
Exposed (1)	A_1	B_1	T_1
Unexposed (0)	A_0	B_0	T_0

Incidence Density Sampling

- Purpose: To obtain a representative sample of person-time at risk in the source population (to assure that **$b_1/b_0=T_1/T_0$**).
- Approach: Select one or more controls from disease-free (at risk) members of the source cohort at the '*instantaneous*' time at which each case occurs.

→ **The probability of control selection is proportinal to the total person-time at risk**

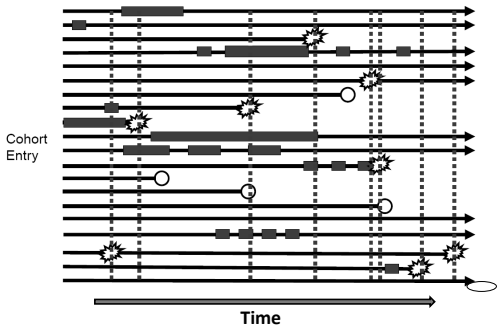
Procedures for Incidence Density Sampling

- Estabish the source cohort and identify cases
- Determine the date on which the first case occurred
- Identify all cohort members (including cases) who were disease free (at risk) at that date (*risk set*)
- Randomly select one (or more) controls from the risk set.
- Repeat steps 1-3 for 2nd, 3rd, ... last case.
- Arrange cases and conrols in a 2x2 table and calculate the case-control odds ratio.

Some Notes on Incidence Density Sampling

- Some selected controls may later be selected as cases, especially if ividence is high. That's OK!
- Some controls may be selected more than once. That's OK, too.
- No need for "rare disease assumtion" in incidence density case-control studies
- Changes in exposure over time are unproblematic (if measurement is unbiased and precise).
- **Probability of a person being selected as a control is proportional to that person's contribution to the total person-time at risk in the source population.**
- **Because sampling from each risk set is random (independent of exposure), the ratio of exposed to unexposed controls (b_1/b_0) is – apart from sampling error– proportional to the ratio of exposed to unexposed person-time in the source cohort (T_1/T_0).**

Incidence Density Sampling



Odds Ratio in Case-Cohort Studies

Recap:

- Average risk exposed = # exposed cases / # exposed at beginning of study **$R_1 = A_1/N_1$**
- Average risk unexposed = # unexposed cases / # unexposed at beginning of study **$R_0 = A_0/T_0$**
- Risk ratio = risk exposed / risk unexposed

Risk Ratio = $(A_1/N_1) / (A_0/N_0)$

In case-control studies, frequency of exposed and unexposed at beginning of study (N_1, N_0) is not observed.

	Disease	Disease-Free	Total
Exposed (1)	A_1	B_1	N_1
Unexposed (0)	A_0	B_0	N_0

Odds Ratio in Case-Cohort Studies

- Risk ratio = risk exposed / risk unexposed

  Risk Ratio = $(A_1/N_1) / (A_0/N_0)$

  can be rearranged to

  Risk Ratio = $(A_1/A_0) / (N_1/N_0)$

= odds of exposure in diseased / ratio of exposed to unexposed at beginning of the study period

In case-control studies, information is not available on all exposed and unexposed subjects at the beginning of the study period (N_1, N_0)… but

If controls (b) are selected to reflect the exposure distribution among the total source population at the beginning of the study period →

$b_1/b_0 = N_1/N_0$

	Disease	Disease-Free	Total
Exposed (1)	A_1	B_1	N_1
Unexposed (0)	A_0	B_0	N_0

Odds Ratio in Case-Cohort Studies

If **RR = $(A_1/A_0) / (N_1/N_0)$**

and **$b_1/b_0 = N_1/N_0$**

we can replace **N_1/N_0** with **b_1/b_0**

→ **Risk Ratio = $(A_1/A_0) / (b_1/b_0)$ … the case-control exposure odds ratio**

If controls (independently of exposure status) are selected from the population at risk at the beginning of the study period, the exposure odds ratio from the case-control study directly estimates the risk ratio in the source population.

	Disease	Disease-Free	Total
Exposed (1)	A_1	B_1	N_1
Unexposed (0)	A_0	B_0	N_0

Odds Ratio in Cumulative Case Control Studies

Recap:

- Incidence odds exposed (cohort) = # exposed cases / # exposed noncases **$O_1 = A_1/B_1$**
- Incidence odds unexposed (cohort) = # unexposed cases / # unexposed noncases **$O_0 = A_0/B_0$**
- Incidence OR = odds disease (exposed) / odds disease (unexposed)

  OR = $(A_1/B_1) / (A_0/B_0)$

In case-control studies, frequency of exposed and unexposed among noncases at the end of the study (B_1, B_0) is not observed.

	Disease	Disease-Free
Exposed (1)	A_1	B_1
Unexposed (0)	A_0	B_0

Odds Ratio in Cumulative Case Control Studies

- Incidence OR = odds disease (exposed) / odds disease (unexposed)

OR = $(A_1/B_1) / (A_0/B_0)$

can be rearranged to

OR = $(A_1/A_0) / (B_1/B_0)$

In case-control studies, frequency of exposed and unexposed among noncases at the end of the study (B_1, B_0) is not observed... but

If controls (b) are selected to reflect the exposure distribution among the noncases in the source populationn at the end of the study period →

$b_1/b_0 = B_1/B_0$

	Disease	Disease-Free
Exposed (1)	A_1	B_1
Unexposed (0)	A_0	B_0

Odds Ratio in Cumulative Case Control Studies

If **OR = $(A_1/A_0) / (B_1/B_0)$**

and **$b_1/b_0 = B_1/B_0$**

we can replace **B_1/B_0** with **b_1/b_0**

→ **OR = $(A_1/A_0) / (b_1/b_0)$ ... the case-control exposure odds ratio**

If controls (independently of exposure status) are selected from the noncases in the source population at the end of the study period, the exposure OR from the case-control study directly estimates the incidence OR in the source population.

Assuming that the disease is rare, the incidence odds ratio approximates the relative risk.

	Disease	Disease-Free
Exposed (1)	A_1	B_1
Unexposed (0)	A_0	B_0

How many controls?

- Sample size refers to precision not to validity!
- Precision increases with the number of controls and cases
- With a fixed number of cases, the proportion of the maximum precision (unlimited controls) that is reached is approximately: **r/(r+1),** where **r** is the ratio of controls to cases

  If r = 4 (4:1 matching) precision is 4/(4+1) = 0.80

  > 4 controls per case of little additional statistical value

How many controls?

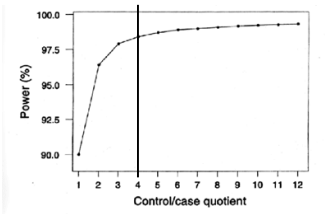


Figure 6.2 Power against control/case quotient. This shows the power to detect an approximate relative risk of 2 when the risk factor has a prevalence of 30%, a two-sided 5% significance test is to be used and 188 cases are available.

Matching in Case-Control Studies

- Can be employed to **increase efficiency**
- Does *not directly control confounding*
- **Introduces selection bias** when the matching variable is a confounder. Because a confounder is by definition associated with exposure, confounder-matched control selection is not independent of exposure.
  → **matching variables have to be controlled for in the analysis**
- Cannot evaluate the matching factor

Bias in Case-Control Studies

- Any issues of confounding and measurement bias present in the source population need to be addressed just like in a cohort study
- Significant ***potential*** for selection bias introduced during the control selection process (control selection directly or indirectly associated with exposure)
- Measurement bias can be more problematic when case/control status is known at the time of measurement (e.g., recall bias when exposure is ascertained directly from the patient)

Common Misconceptions

- Case control studies are inferior to cohort studies
- Rare disease assumption necessary for all case-control studies
- Measurement is inherently worse in case-control studies (recall bias)

Take-Home Messages

- Case-control studies are a widely used and powerful research method
- Think of case-control studies as more efficient versions of the corresponding cohort studies
- Particularly, case-control studies are much more efficient than cohort studies when outcomes are rare and when exposure or confounder information is expensive to collect (e.g., genotyping)
- Validity depends upon whether controls provide a clear view of the population from which the cases arise

Take-Home Messages (cont)

- Depending on the sampling paradigm used to select the controls, the case-control odds ratio estimates different measures of association
- The odds ratio from incidence density case-control studies is equivalent to the rate ratio in the underlying cohort (no rare disease assumption necessary)
- Confounding and other biases need to be addressed (just as in cohort studies)
- Selection bias and generalization can be problematic, especially in studies with secondary bases.

Thank you for your attention!

Contact:
tgerhard@rci.rutgers.edu