School of Optometry

Case-Control Studies

Dr. John Waterbor
UAB School of Public Health
October 25, 2010

Outline

How a Case-Control Study is Done
Purpose
Subjects for Study
Choosing cases
Choosing controls
Population-based vs. hospital-based controls
Control selection issues
Analysis of Data
Odds Ratio and Its Interpretation
Evaluation of dose-response
Strengths and Limitations
Nested Case-Control Studies

- *Reference*: "The Evolving Case-Control Study" by Philip Cole (now Professor Emeritus at UAB)

How a Case – Control Study is Done

- The ***case-control study design*** is a very common ***observational*** study design used in epidemiology
- In brief, the ***history*** of exposure to a putative (suspected) cause of a disease is compared between people who have that disease (***cases***) and similar people who ***do not*** have that disease (***controls***).
- If the ***cases*** have ***substantially more*** exposure history than do the ***controls***, one possible explanation is that the exposure ***caused*** the disease that the cases have (but other explanations for the difference in exposure history are possible).

Purpose of the Case-Control Design

- The purpose of the ***case-control*** study design is to investigate ***associations*** between exposures and diseases so that disease ***etiology*** can be better understood.
- The investigator is in search of ***the truth*** in the population about the exposure – disease relationship.
- Note that another term for "case-control study" is "***retrospective study***". . . but do not confuse this design with the "retrospective ***follow-up***" design !

Basis for Choosing Subjects (Cases and Controls)

- ***Note that cases*** and ***controls*** are ***chosen*** based on their ***disease status*** (***D***+ or ***D***-) without our knowing their exposure status (***E***+ or ***E***-).
- If you chose only ***cases who were exposed*** and ***controls who were unexposed***, your study would be ***biased away from the null*** (in fact, to "infinity") – so, choose cases and controls "***blindly***" with respect to their exposure to ***E***.

Identification and Selection of Cases

- People who have a disease ***D*** and who are selected for study are called the ***cases (D+)***.
- ***Prevalent*** cases may be chosen (i.e., subjects having ***D*** without regard to recency of their diagnosis) but if their diagnosis was long ago, it may be difficult to verify whether the exposures they report were experienced ***prior*** to or ***after*** diagnosis (ref: ***Hill's Criteria***).
- ***Incident cases*** are generally preferable! Enroll them in the study ***soon*** after diagnosis so that ***all*** exposures they report must have occurred ***prior*** to diagnosis.

Source of Cases

- ***Cases*** may be drawn from a ***population*** (geographically defined, as from a state cancer registry; or from the rolls of a health insurance plan, or from another large database) . . . or the cases may be drawn from a ***hospital*** (based on admission diagnosis or discharge diagnosis)
- The ***case group*** is then said to be ***"population-based"*** or ***"hospital-based."***
- The ***base*** of the ***control group*** should match the ***base*** of the ***case group*** (***population*** vs. ***hospital***)

Identification and Selection of Controls

- Other people who are ***comparable*** to the ***cases demographically*** and on important potential ***confounders*** but who ***do not*** have the disease are selected to be the ***controls (D-).***
- ***A common misconception is that the controls should be representative of the general population! This is untrue – controls should be comparable to the cases with regard to demographics and possible confounders.***

"Rules of Thumb" for Choosing Controls

The choice of the ***controls*** must be done with care because use of a "bad" control group can yield misleading results! Here are some ***principles of control selection:***

1) ***Controls*** should be drawn from the same ***population*** (or, from the same ***hospital***) from which the ***cases*** were drawn.
2) ***Generally, a control could have, or would have, become a case*** in your study if he or she had been diagnosed with ***D*** during the study period.
3) ***Controls*** should have had the same ***exposure opportunity*** as did the ***cases.***
4) ***Do not choose controls whose exposure history is known*** with certainty before the data collection begins! This would make your entire study meaningless!

Population-based vs. Hospital-based Controls

In general, ***controls*** are drawn from the ***same source*** that provided the ***cases***: either from a "***population***" (e.g., a geographic population, members of a health insurance plan, etc.) or from one or more ***hospitals*** (or clinics).

Then the ***control group*** is said to be "***population-based***" or "***hospital-based***."

Do not mix sources of ***controls:*** draw ***all*** of your controls from the ***population*** . . . or draw ***all*** of your controls from one (or more) ***hospitals*** that you have chosen.

Population-based Controls

Strength: their exposure history reflects the exposure history of the ***population***, which is desirable if the ***cases*** were drawn from that population.

Limitation: they may have little interest in being studied so the quality and quantity of ***information*** they provide and their ***cooperation*** may be ***poor***.

Hospital-based Controls

Strength: they tend to be ***cooperative*** and full of detailed information because they are a "captive audience" that may want to help future patients.

Limitation: the ***validity*** of the study may suffer because hospitalized people have diseases that may be caused by the ***same exposures*** (e.g., cigarettes, alcohol, inactivity, bad diet, compromised immunity, etc.) that we are trying to study. Would the use of ***controls*** who have substantial ***exposure histories*** drive the point estimate of the ***RR toward*** the null or ***away*** from it?

A Few Examples of Unresolved Control Selection Issues

Within a single study is it okay to use ***two or more*** different control groups? If you do this and the results differ which results would you believe?

Some epidemiologists say that ***hospital*** controls should ***never*** be used.

In a study of decedents (***dead*** cases), should you make a point to choose ***dead*** controls?

If you need to interview ***surrogates*** (relatives or friends) of stroke ***cases*** who are unable to communicate, should you also interview ***surrogates*** of the ***controls*** even though these controls ***can*** communicate?

Basic Data Collection and Analysis

- Information on the ***history of exposure*** to possible ***causes*** of disease ***D*** in the ***period of time prior to diagnosis of the cases*** (ref: ***Hill's Criteria***), is gathered from the ***cases*** and from the ***controls*** by means of questionnaires, interviews, or abstraction of medical or occupational records.
- A ***Relative Risk*** is calculated. If an ***association*** (***non-null*** result) is found, this is evidence (but not proof!) of ***causation*** (or, evidence of ***prevention*** if the association is ***inverse***).

The Analysis of Case-Control Data

The analysis of data from a case-control study in its simplest form is by means of a ***2 x 2 table*** where the ***exposure histories*** of ***cases*** and ***controls*** are ***compared***.

As is true for other study designs, the basics of the analytic results of ***case-control*** studies are:

1) ***Point Estimate of the Relative Risk*** – using a new variety of ***RR***, the ***odds ratio*** ("***OR***")
2) ***Precision of the point estimate*** – a ***confidence interval*** or a ***p value*** (see next lecture).

The Odds Ratio as the Measure of Association (Relative Risk)

The ***measure of association*** that we calculate in a ***case-control*** study is the ***Odds Ratio***.

Like other ***RR*** measures the ***OR*** has:

Range: (0 , ∞) ("unbounded")

Null value: 1

Explanation of the Odds Ratio

The ***odds ratio*** is the ***ratio*** of two ***odds***:

OR = Odds of exposure history among cases
Odds of exposure history among controls

Where the "***odds of exposure history***" is
P (E+) / P (E-)

Note that an "***odds***" is simply the ***ratio of two probabilities***: the probability of "***yes***" divided by the probability of "***no***."

Layout of 2 x 2 Table for Case-Control Data and Computation of the OR

	CA	CO	
E+	a	b	a+b
E-	c	d	c+d
	a+c	b+d	n

OR = (a / a + c) / (c / a + c) / (b / b + d) / (d / b + d)

OR = (a / c) / (b / d) = ad / bc

So, in practice the ***OR*** can be calculated as the ***"cross-product" ratio:*** divide the ***product*** of the ***concordant cells*** (E+ and D+; E- and D-) by the ***product*** of the ***discordant cells*** (E+ and D-; E- and D+).

Example of Layout of Case-Control Data in a 2 x 2 Table

	CA	CO	
E+	30	10	40
E-	60	70	130
	90	80	170

Start with ***90 cases*** and ***80 controls*** (***170 subjects***).

Then we find that exposure history (***E+***) is reported by ***30 cases*** and ***10 controls***. This leaves ***60 cases*** and ***70 controls*** in the unexposed (***E-***) category.

Computation of the Point Estimate of the Odds Ratio

OR = [30 (70) / 10 (60)] = 21 / 6 = 3.5

Literal interpretation*:** the ***odds of exposure history is ***3.5 times greater*** among ***cases*** than among ***controls***.

One ***possible explanation*** for this finding is that the exposure in fact ***caused*** the disease: that is why a higher proportion of cases have exposure history!

We should also calculate ***95% confidence limits*** (or a ***p value***) as a measure of ***precision*** of this point estimate.

A Useful Interpretation of the OR if Incident Cases Were Used

If the ***cases*** were population-based ***incident*** cases (as opposed to ***prevalent*** cases), it can be shown that the ***OR = IR***.

Under these circumstances we can ***interpret*** the ***OR*** using "follow-up study" language which may be more understandable:

The disease is 3.5 times more common among the exposed than among the unexposed.

Evaluation of Dose-Response

Dose-response can be evaluated in a ***case-control*** study if exposure history ***E*** is measured on a ***continuous*** scale or is otherwise ***categorized*** into ***3 or more levels***.

If there is a positive association between ***E*** and ***D***, the ***cases*** will have a distribution of exposure history that is weighted toward higher exposures than is the distribution among the ***controls.***

A "string" of ***OR***'s according to ***E*** level can be calculated and then the ***Mantel Extension Test for Trend*** can be done.

Strengths of the Case-Control Design

1) ***Well-suited to the study of rare diseases.*** Even if we know of only a few cases, controls can be matched to those cases and the study can proceed.
2) ***Well-suited to the study of diseases having long induction periods*** because the induction period for the cases has already passed by the time they are studied. Although . . . if the induction period is ***extremely long*** some subjects may have difficulty ***remembering*** when or whether they were exposed.
3) ***Usually efficient*** in terms of the amount of ***time*** needed to conduct the study and ***cost*** of the study.

Limitations of the Case-Control Design

1) ***Incidence rates and cumulative incidences of D cannot be calculated*** because the ***sizes*** of the source populations that produced the cases and the controls are unknown,. (You can think of the ***cases*** and the ***controls*** as the "***numerators***" of ***D+*** and ***D-*** people, respectively, drawn from populations whose "***denominators***," are unknown.)
2) ***Selection bias*** can be problematic because the choice of an inappropriate or controversial ***control group*** can call your results into question and in fact can produce misleading results (a ***"biased" RR***).
3) ***Information bias*** can be problematic if ***cases*** tend to over-estimate (or perhaps, under-estimate) their exposure histories. ***Controls*** tend to relate true exposure histories or sometimes underestimate them.

Example of Cases Over-reporting Their Exposure Histories

- In a ***case-control*** study of smoking and lung cancer, newly diagnosed lung cancer ***cases*** are interviewed about their smoking histories. Because they "know" that smoking caused their cancer or because they feel guilty about it, they may ***over-estimate*** or ***embellish*** the report of their smoking history.
- If the ***controls*** (who ***do not*** have lung cancer and who have no "vested interest" in the study) report their smoking history accurately, does this create a bias ***toward*** or ***away from*** the null? (More on ***bias*** later!).

Contrasting Approaches for Studying Cigarette Smoking and Lung Cancer

- ***Follow-up*** design approach: identify ***E+*** and ***E-*** people (i.e., ***smokers*** and ***non-smokers***) and follow them to determine whether these groups differ in their ***incidence of lung cancer***.
- ***Case-control*** design approach: identify ***D+*** and ***D-*** people (i.e., ***lung cancer cases*** and ***controls***) and by means of interviews or record abstraction determine whether these groups differ in the intensity and duration of their ***cigarette smoking history***.

Calculation of Attributable Proportion in Case-Control Studies

In the formulas below, ***OR*** = odds ratio,

CA_E = Proportion of ***Cases*** exposed and

CO_E = Proportion of ***Controls*** exposed.

The ***Attributable Proportion among the Exposed*** is

$AP_E = (OR - 1) / OR$ (as you might expect!).

There are 3 equivalent formulas for the ***Attributable Proportion in the Population***:

$AP_{POP} = AP_E (CA_E) = (CA_E - CO_E) / (1 - CO_E) = [CO_E (OR - 1)] / [CO_E (OR - 1)] + 1$

Example of AP_E and AP_{POP} Calculations for Case – Control Data

	CA	CO
E+	60	100
E-	40	100
	100	200

$AP_E = [(60 \times 100) / (40 \times 100)] - 1 / [\ ]$
$= (1.5 - 1) / 1.5 = 0.5 / 1.5 = 1 / 3 = 33\%$

(20 of the ***60 exposed*** cases are "due to" the exposure)

$AP_{POP} = AP_E\ (CA_E) = (1 / 3)\ (60 / 100) = 20\%$

(or try the other formulas to get the same answer!)

(20 of the ***100 total*** cases are "due to" the exposure)

Nested Case-Control Studies

A ***nested*** case-control study is a ***variant*** of the case-control study design.

The ***nested*** case-control study chooses its cases and its controls from the ranks of a ***cohort*** observed in a ***follow-up study***.

From the ***follow-up study records***, choose some or all of those individuals who developed the disease to serve as ***cases***, and choose some or all of those individuals who ***did not*** develop the disease to serve as ***controls***, for your ***nested case-control*** study.

Why a Nested Case-Control Study Would be Done

Why would a ***nested case-control*** study be done if a ***follow-up study*** has already been done?

1) To explore ***specific aspects*** of an exposure found to be associated with a disease in the follow-up study. (See Example 1).
2) To evaluate an ***unexpected outcome*** that may have been caused by an exposure ***not assessed*** in the follow-up study. (See Example 2).

Nested Case-Control, Example 1

Example 1: In a ***follow-up*** study of ***occupational groups*** and ***cancer*** it is found that female ***school teachers*** have an elevated rate of ***breast cancer***. Which ***specific aspects*** of being a teacher could explain this finding?

From the cohort, identify the female school teachers who were diagnosed with breast cancer (***cases***) and some or all of the female school teachers who were ***not*** diagnosed with breast cancer (***controls***). Collect data on history of exposure to ***specific*** factors that you believe may be causal (e.g., using the copy machine? eating in the school cafeteria?) and estimate the ***OR*** in a case-control analysis.

Nested Case-Control, Example 2

Example 2: In a ***follow-up*** study of ***sunlight exposure*** and development of ***melanoma*** among young adults, ***unexpectedly*** many subjects are found to develop ***lung cancer***.

A ***nested*** case-control study could be conducted to assess the role of ***smoking***, which was ***not even measured*** in the follow-up study because lung cancer was not an expected outcome or even an outcome of interest. Other hypothesized causal exposures could also be investigated.

--JW (caseco08)