Case-control studies in epidemiological research

Samiran Sinha
Texas A&M University

Nov 18, 2021

Case control study

- In a case-control study design two groups of subjects, one with the disease (or a health outcome) and the other without the disease are included. Exposure information are collected from these subjects. The subjects with the disease are referred to as cases while the subjects without the disease are referred to as controls.
- The cases and controls are not selected randomly from the underlying population. Rather, case subjects are, generally, randomly selected from the population of diseased people and controls are randomly selected from the population of disease-free people. Importantly, the sampling must not depend on subjects' exposure status.
- The purpose of the case-control study is to find the connection/association/effect of exposure on the disease.

Advantages/disadvantages

- The design is highly efficient, it can be done in a much shorter time period (less costly) compared to a similar cohort study.
- The design can be used to study rare and common diseases.
- If an exposure is rare in the population, then whether we use a case-control or cohort study, both designs need a large sample size. However, if the exposure is strongly associated with the disease, then a case-control study will be much more efficient than a cohort study.
- In a case-control design, cases and controls must be comparable, otherwise results are biased.
- Since the exposure information is collected retrospectively, the measurements are subject to recall bias that may lead to bias results.

Case control study

- Suppose that we have a case-control data with a scalar dichotomous exposure variable, call it X. Let Y be the disease variable taking on zero and one for control and case subjects, respectively.
- The data can then be summarized as follows:

	Unexposed ($X = 0$)	Exposed ($X = 1$)	
Control	a	b	n_0
Case	c	d	n_1

n_0 : total number of controls
n_1 : total number of cases

Case control study

Two common measures of associations: relative risk (RR) and odds ratio (OR)

- $RR = \text{pr}(Y = 1|X = 1)/\text{pr}(Y = 1|X = 0)$, a ratio of the risk of the disease from exposed to unexposed, generally RR is not estimable from case-control data
- $$\begin{aligned} OR &= \frac{\text{Odds of the disease among exposed}}{\text{Odds of the disease among unexposed}} \\ &= \frac{\text{Odds of exposure among disease}}{\text{Odds of the exposure among non-disease}} \end{aligned}$$
- Fortunately, OR is estimable from a case-control data, and its estimator is

$$\widehat{OR} = \frac{ad}{bc}.$$

Case control study

- When the exposure and disease are independent, $RR = 1$ and $OR = 1$.
- In the presence of dependence, $RR \neq 1$ and $OR \neq 1$.
- Thus, testing of independence is equivalent to test $OR = 1$ or $\log(OR) = 0$.
- $OR > 1$: Odds of the disease is greater in the exposed group than in the unexposed group, exposure is a risk factor
- $OR < 1$: Odds of the disease is less in the exposed group than in the unexposed group, exposure is a protective factor

Case control study, a toy example

A dataset on average daily consumption of alcohol and oesophageal cancer [1]

	0-79 g ($X = 0$)	80+ g ($X = 1$)	
Control	666	109	775
Case	104	96	200

- $\widehat{OR} = 666 \times 96/104 \times 109 = 5.64$
- The standard error of $\widehat{\log\text{-OR}}$ is $\sqrt{666^{-1} + 109^{-1} + 104^{-1} + 96^{-1}} = 0.17$.
- The 95% large sample CI of the OR is $\exp\{\log(5.64) \pm 1.96 \times 0.17\} = (3.99, 7.92)$.

[1] *Statistical Methods in Cancer Research* by Breslow and Day, 1980, p. 123.

R code

Code

```
# Uploading the necessary package
library(epitools)

# Entering the data
 data=array(c(666, 104, 109, 96), dim=c(2, 2),
 dimnames=list(y=c(0, 1), x=c(0, 1)))
> data
   x
y     0   1
  0 666 109
  1 104  96
```

R code

Code

```
> oddsratio(data)
$data
       x
y         0   1 Total
  0     666 109   775
  1     104  96   200
  Total 770 205   975
$measure
   odds ratio with 95% C.I.
y   estimate    lower    upper
  0 1.000000       NA       NA
  1 5.624773 3.992007 7.947159
$p.value
   two-sided
y   midp.exact fisher.exact   chi.square
  0         NA            NA           NA
  1          0 1.079486e-22 8.614569e-26
$correction
[1] FALSE
```

Interpretation

- The odds of the disease (or cancer) among the exposed group is 5.64 times that of among the unexposed group.
- If we know, the disease is rare (less than 1% in the population), then the odds ratio can be interpreted as the relative risk. Then we can say, the risk of the disease among exposed group is 5.64 times that among the unexposed group.
- We do not claim that (3.99, 7.92) includes the true OR with 0.95 probability. Rather, in repeated sampling from the underlying population and computation of the 95% CI, about 95% of the intervals include the true OR. However, we do not know if (3.99, 7.92) includes the true OR nor we can associate any probability with this interval. Just say that we are 95% confident that the true odds ratio is in (3.99, 7.92).

- The above example indicates the risk of the esophageal cancer among exposed group is 5.64 times that among the unexposed group.
- Let's look at this more closely. Let $\mathrm{pr}(Y = 1|X = 1) = 0.00056$ and $\mathrm{pr}(Y = 1|X = 0) = 0.0001$, and the relative risk in this case 5.6. In another situation, $\mathrm{pr}(Y = 1|X = 1) = 0.28$ and $\mathrm{pr}(Y = 1|X = 0) = 0.05$, this also results in the relative risk of 5.6. In one case, the expected number of cancer cases increases from from 1 to 5.6 when 10,000 unexposed cases all become exposed. In the second case, the expected number of increase is from 5 to 28 when 100 unexposed subjects all become exposed.
- Obviously, assuming that the risk is a causal, the second case is more of a public health concern than the first case.
- However, only the relative risk is not able to shade light on this entire situation– particularly, it does not tell us the gravity of the problem in the view of public health.

An alternative measure

Population attributable risk (PAR) measures **how much of the disease in the population is caused by the risk factor**. PAR is measured as the excess rate of disease in the total study population of exposed and unexposed individuals compared to the population of unexposed individuals, and is formulated as

$$\text{PAR} = \frac{\text{Risk of the disease in total population} - \text{Risk of the disease among the unexposed}}{\text{Risk of the disease in total population}}.$$

In terms of notations,

$$\text{PAR} = \frac{\text{pr}(Y = 1) - \text{pr}(Y = 1|X = 0)}{\text{pr}(Y = 1)}.$$

- It can also be expressed as

$$\mathrm{PAR} = \frac{P_e(R-1)}{P_e(R-1)+1},$$

  where P_e stands for the proportion of population exposed and R stands for the relative risk of the disease from exposed to unexposed.

- For the esophageal cancer, we may assume that **1)** the control population is very similar to the general population, and **2)** controls are sampled randomly from the populations. Then P_e can be estimated by the proportion of exposed in the control sample, 0.14. Thus, $\widehat{\mathrm{PAR}} = 0.14(5.6-1)/\{0.14(5.6-1)+1\} = 0.39$.

- Thus, about 39% of the disease in the general population is caused by the exposure (average daily alcohol consumption 80+ g).

- Although simple, the analysis of a 2×2 table is a over simplification, and may risk a confounding bias. For our toy example, the age effect likely be confounded in the disease-exposure association.
- Suppose now that the data are stratified for different age groups. It is presented in the next slide.

Several 2×2 tables

Age		0-79 g ($X = 0$)	80+ g ($X = 1$)	$\widehat{\text{OR}}$
25-34	Control	106	9	∞
	Case	0	1	
35-44	Control	164	26	5.05
	Case	5	4	
45-54	Control	138	29	5.67
	Case	21	25	
55-64	Control	139	27	6.36
	Case	34	42	
65-74	Control	88	18	2.58
	Case	36	19	
75+	Control	31	0	∞
	Case	8	5	

Questions to ask ourselves

- It is clear that the estimated conditional odds ratios are changing. Are the underlying true conditional odds ratios different?
- To check this, test the homogeneity of odds ratios (Breslow-Day test). Let θ_k be the conditional odds ratio for age group k.
- Set $H_0 : \theta_1 = \cdots = \theta_6$ versus H_a : at least one of θ_k's is different from the rest.

R code

Code

```
# Uploading the necessary package
library(DescTools)
# Data input
data2=array(c(106, 0, 9, 1, 164, 5, 26, 4, 138, 21, 29, 25, 139, 34, 27,
42, 88, 36, 18, 19, 31, 8, 0, 5), dim=c(2, 2, 6), dimnames=list(y=c(0, 1),
x=c(0, 1), age=c("25-34", "35-44", "45-54", "55-64", "65-74", "75+")))

BreslowDayTest(data2)

Breslow-Day test on Homogeneity of Odds Ratios

data:  data2
X-squared = 9.3234, df = 5, p-value = 0.09684
```

Breslow-Day test result

- Since the p-value is large, we fail to reject H_0 at the 5% level, and conclude that there is not enough evidence that the underlying conditional odds ratios are different (seems contradictory, isn't it?).
- Important point is that if cell frequencies are small, then Breslow-Day (BD) test is not very reliable. Then one should fit a logistic regression model and then carry out the test.
- Since BD test was non-significant at the 5% level, we now aim to estimate that common conditional odds ratio. This odds ratio is adjusted for age.

Common odds ratio adjusted for the confounding variable

Mantel-Haenszel estimate:

$$\widehat{\theta}_{mh} = \frac{\sum_i a_i d_i / N_i}{\sum_i b_i c_i / N_i}$$

- $+$: It is not affected by zero cell frequencies
- $-$: The standard calculation is very tedious
- Next slide we show how to obtain the MH estimate and conduct the hypothesis test of conditional independence of the disease and the exposure.

R code

Code

```
# Uploading the necessary package
library(stats)

mantelhaen.test(data2)

Mantel-Haenszel chi-squared test with continuity correction

data:  data2
Mantel-Haenszel X-squared = 83.215, df = 1, p-value < 2.2e-16
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 3.562131 7.467743
sample estimates:
common odds ratio
         5.157623
```

MH test results

- The age adjusted conditional odds ratio is 5.2 with a 95% CI (3.6, 7.5).
- With the p-value smaller than 0.05, we reject H_0, and conclude that there is a strong conditional association between the disease and heavy alcohol consumption.

Some comments[2]

- When the conditional odds ratios are the same across the strata, and they are equal to the crude odds ratio, the potential confounding variables do not have any confounding effect.
- When the conditional odds ratios are the same across the strata, and they are not equal to the crude odds ratio, the potential confounding variables carry confounding effect.
  - If the common conditional odds ratio is larger than the crude odds ratio, the confounder is termed as negative confounder.
  - If the common conditional odds ratio is smaller than the crude odds ratio, the confounder is termed as positive confounder.
- When the conditional odds ratios are not the same across the strata, we call the confounder as an effect modifier.

[2] M.C. Costanza, Matching, *Preventive Medicine*. 1995; 24:425–433.

Logistic regression

Why?

- For very small cell frequencies, BD test for testing homogeneity of odds ratio is not reliable.
- The number of confounding variables may not be a scalar or categorical variable. In the presence of a large number of confounding variables, MH method or BD test is not practical.

These issues can be overcome by fitting a logistic model. Of course, it requires some parametric model assumptions.

Logistic regression

- High level of daily alcohol consumption: $X = 1$, not high level of alcohol consumption: $X = 0$
- Age: confounding variable (Z), treat it as a numeric with $Z = 1$ for age 25-34, $Z = 2$ for age 35-44 etc.
- Model π, the probability of $Y = 1$ given X and Z:

$$\text{logit}(\pi) = \alpha_0 + \alpha_1 X + \alpha_2 Z + \alpha_3 XZ$$

- In the above model the age effect is assumed to be linear (a parametric assumption).

- If the interaction term α_3 is zero, then the conditional odds ratios are homogeneous. So, test of homogeneity can be done by testing $H_0 : \alpha_3 = 0$ versus $H_a : \alpha_3 \neq 0$.
- If the above test is statistically significant, then it is meaningless to estimate the age adjusted common odds ratio (because there is no such common odds ratio). Rather, we should report the conditional odds ratio for each age group. Also, the confounding variable is termed as an effect modifier.

- If the above test is statistically non-significant, then we can report the age adjusted common odds ratio, $\exp(\alpha_1)$.
- We can make statistical inference on this conditional odds ratio.
- The logistic model can include many confounders.

R code

Code

```
# Preparation of the data in the desirable format
mydata=as.data.frame.table(data2)
disease= mydata$Freq[seq(2, 24, 2)]
total=mydata$Freq[seq(1, 24, 2)]+ mydata$Freq[seq(2, 24, 2)]
exposure= mydata$x[seq(2, 24, 2)]
age= as.numeric(mydata$age[seq(2, 24, 2)])

# Invoking the GLM function to fit the logistic model
out=glm(disease/total~exposure+age+exposure*age, family=binomial,
weight=total)
 summary(out)
```

R code

Code

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.58762  -1.72026   0.08124   1.19703   1.49961

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.09127    0.36114 -11.329  < 2e-16 ***
exposure1     1.75105    0.63839   2.743  0.00609 **
age           0.61368    0.08531   7.193 6.32e-13 ***
exposure1:age 0.00779    0.16424   0.047  0.96217
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 211.608  on 11  degrees of freedom
Residual deviance:  31.929  on  8  degrees of freedom
AIC: 80.257
```

Logistic regression

- The interaction term is statistically non-significant at the 5% level.
- The age adjusted common odds ratio estimate of the disease-exposure association is $\exp(1.75) = 5.75$ with a 95% CI $\exp(1.75 \pm 1.96 \times 0.64) = (1.64, 20.17)$.
- Why does this OR estimate differ from the MH estimate?
- For a case-control data, the intercept term has no useful interpretation.
- For any given exposure group, the odds ratio of the disease for 10 years increase in age is $\exp(0.61) = 1.84$. In other words, for any given exposure group, the odds of the disease increases by 84% for 10 years increase in age.

Further control on confounding

- As a strategy to control confounding, often matched data are analyzed. Matching can be introduced at the design stage or at the data analysis stage. In essence, for every case, a number of controls are chosen/included in the study by matching the values of a set of potential confounding variables.
- If the matching is introduced at the design stage, then we can keep a constant ratio of cases to controls in each stratum defined by the matching variables. If the matching is done before the analysis (not at the data collection stage), then there is a possibility of imbalance in the ratio of cases to controls across the strata.
- Two basic objectives of matching: better control of the confounding effect and increase the efficiency of measuring the strength of the association
- For the analysis of matched case-control data, conditional logistic regression method is used.

A toy example of matched data [3]

Code

```
library(epiDisplay)
 data(VC1to6)
> head(VC1to6)
  matset case smoking rubber alcohol
1      1    1       1      0       0
2      1    0       1      0       0
3      2    1       1      0       1
4      2    0       1      1       0
5      3    1       1      1       0
6      3    0       1      1       0
```

Matching variables: age, sex, neighborhood

rubber: worked in the rubber industry

[3] Chongsuvivatwong, V. 1990 A case-control study of esophageal cancer in Southern Thailand. J Gastro Hep 5:391–394.

R code for the conditional logistic analysis

Code

```
library(survival)
 out5=clogit(case~smoking+rubber+alcohol+strata(matset), data=VC1to6)
> summary(out5)
Call:
coxph(formula = Surv(rep(1, 119L), case) ~ smoking + rubber +
    alcohol + strata(matset), data = VC1to6, method = "exact")

  n= 119, number of events= 26
         exp(coef) exp(-coef) lower .95 upper .95
smoking     1.5523     0.6442    0.4375     5.508
rubber      0.6331     1.5797    0.1780     2.251
alcohol     5.2951     0.1889    1.6490    17.003

Concordance= 0.688  (se = 0.065 )
Likelihood ratio test= 12  on 3 df,   p=0.007
Wald test            = 9.18  on 3 df,   p=0.03
Score (logrank) test = 11.24  on 3 df,   p=0.01
```

Further control on confounding

- After adjusting the effect for the confounding variables (age, sex, neighborhood), and other risk factors, smoking and rubber, the odds ratio of association between cancer and alcohol is 5.3.
- Note that smoking and rubber are not confounding variable, they are just potential co-risk factors. The data were also analyzed by including `alcohol` $\times$ `smoking` and `alcohol` $\times$ `rubber` interaction terms, and both terms turned out statistically non-significant. This indicates that they are not effect modifiers.

Remarks

- Researchers are worried that matching and conditional analysis does not reduce the confounding bias always. Sometimes, matching may introduce bias.[4]
- If all potential confounding variables are measured, then matched data can be used to estimate the causal odds ratio of association.[5]
- Observational studies need not be cohort or case-control studies. There are many variations, and among them nested case-control study (case-controls are nested within a cohort) is popular choice. It is efficient, and provides more information on the disease etiology, such as the incidence rate of the disease with the chronological age of the subjects and its association with the risk factors, compared to a simple case-control study.[6]

[4] M.C. Costanza, Matching, *Preventive Medicine*. 1995; 24:425–433.

[5] Rose S, Laan MJ. Why match? Investigating matched case-control study designs with causal effect estimation. *Int J Biostat*. 2009;5(1):1

[6] V.L. Ernster, Nested Case-Control Studies, *Preventive Medicine*. 1994; 23:587–590.

Further references

Missing exposure data in matched case-control studies:

- Sinha, S. (2010). An estimated-score approach for dealing with missing covariate data in matched case-control studies. *Canadian Journal of Statistics*.
- Sinha, S., and Wang, S. (2009). A new semiparametric procedure for matched case-control studies with missing covariates. *Journal of Nonparametric Statistics*.
- Sinha, S. and Maiti, T. (2007). Analysis of matched case-control data in presence of nonignorable missing data. *Biometrics*.
- Sinha, S., Mukherjee, B., Ghosh, M., Mallick, B. K., Carroll, R. J. (2005). Semiparametric Bayesian analysis of matched case-control studies with missing exposure. *Journal of the American Statistical Association*.

Further references

Multiple disease states in matched case-control studies:

- Ahn, J., Mukherjee, B., Gruber, S. B., and Sinha, S. (2011). Missing exposure data in stereotype regression model: application to matched case-control study with disease subclassification. *Biometrics*.

  Mukherjee, B., Liu, I., and Sinha, S. (2007), Analyzing matched case-control data with multiple ordered disease states, possible choices, and comparisons. *Statistics in Medicine*.

- Sinha, S., Mukherjee, B., and Ghosh, M. (2004). Bayesian semiparametric modeling for matched case-control studies with multiple disease states. *Biometrics*.

Further references

- Gene-environment interaction: Mukherjee, B., Zhang, L., Ghosh, M., and Sinha, S. (2007). Semiparameteric Bayesian analysis of case-control data under gene-environment independence and population stratification. *Biometrics*.
- Sample-size calculation: Sinha, S., and Mukherjee, B. (2006). A score test for determining sample size in matched case-control studies with categorical exposure. *Biometrical Journal*.

Thank you!