Statistical methods in public health
Case-control (CC) studies

Tommi Härkänen

National Institute for Health and Welfare (THL)
Department of Health (TERO)

October 13, 2015

Contents

Keogh RH, Cox DR (2014). *Case-Control Studies*. Cambridge University Press.

Retrospective study design

Prospective designs In e.g. cohort studies we first observe the (potential) exposures (risk factors), and the outcome later.

Retrospective designs In case-control (CC) studies we first observe the outcome, and after that we collect information about the (potential) exposures.
E.g. select all new disease cases in a (sub)population during a time period, and after that select appropriate controls for the cases.
"Do the cases have a higher prevalence of the exposure than the controls?"

Benefits of retrospective designs:

Rare outcomes If a disease is rare, it analyses of the effects would require a large data set in order to observe sufficient number of disease cases.

Short time period to collect data There is little need to wait for disease cases to occur

Data from a unmatched CC study

Binary exposure and binary outcome

		Controls $Y=0$	Cases $Y=1$
Unexposed	$X=0$	$n_0 - r_0$	$n_1 - r_1$
Exposed	$X=1$	r_0	r_1

We want to estimate the prospective association e.g. $\mathbb{P}\{Y \mid X\}$. Can we estimate it using the retrospective design?

Positive exposure? Odds among cases is $r_1/(n_1 - r_1)$ and among controls $r_0/(n_0 - r_0)$. The odds ratio (OR) is

$$\frac{r_1/(n_1 - r_1)}{r_0/(n_0 - r_0)}. \tag{1}$$

Positive outcome? Odds of case vs. control among exposed is r_1/r_0 and among unexposed $(n_1 - r_1)/(n_o - r_0)$. The OR equals (1):

$$\frac{r_1/r_0}{(n_1 - r_1)/(n_o - r_0)} = \frac{r_1/(n_1 - r_1)}{r_0/(n_0 - r_0)}.$$

Selecting cases

Cases can be

Incident cases Only new cases during a time interval are selected.

Prevalent cases All individuals who had the outcome before some time point. (Less reliable due to possible selection mechanisms, rarely used.)

Select cases based on

Population based (primary base)

- Geographical area
- Time interval

"Convenient" source not based on a clearly defined population (secondary study base)

- Hospital

Selecting controls

Population based Controls can be easily selected using a (stratified/weighted/...) random sample (in principle).

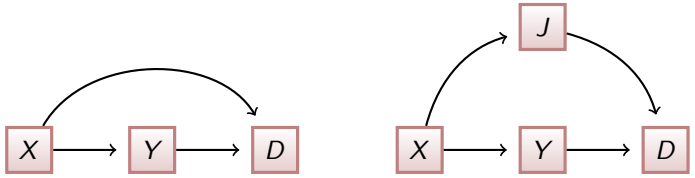
Secondary based Population not defined, so background of controls easily differs from that of cases. E.g. if cases from a hospital, then controls

- other patients from same hospital or
- healthy individuals from same town/city/country?

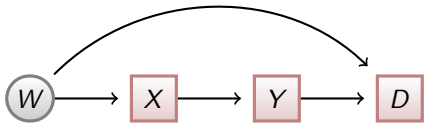
Problem: Latent confounders or other background factors (as generally in observational studies).

Selection bias

Selection D depends not only on the outcome Y but also on exposure X in an unknown way (possibly via other variables J):



(Unobserved) background variables W can also cause selection bias:



Bias due to retrospective exposure ascertainment

Recall bias Cases might remember expose history better ot more selectively than controls.

Information bias Interviewer might be influence by the outcome (e.g. diagnosis) of the study subject. (Do blinding if possible!)

Outcome affects (biological) measurements exposure values can be quite different after the outcome than before.

Some notions:

- If the exposure has been measured prospectively (e.g. baseline measurement of a cohort study), but cases and controls retrospectively during the follow-up, these problems can be avoided.
- In retrospective designs *measurement errors* can be more common and differential between cases and controls.
- Items 1 and 3 are especially common in analyzing prevalent cases.

Confounding

CC studies are observational studies, thus the results are subject to confounding. The issue of adjustment in CC studies is more complex than in prospective studies:

Confounding variables Variables that affect both the exposure X and outcome Y.

Background or intrinsic variables These variables do not affect X but they affect outcome Y

Methods to handle confounding in CC studies:

Selection of controls By selecting controls similar to the cases (w.r.t. background factors and confounders) the need for adjustments can be reduced. E.g. frequency sampling and individual matching.

Adjusted analyses Common methods are stratified analyses and pooling of the results, and regression methods.

Unadjusted and adjusted estimates of log OR

Unadjusted estimator is logarithm of (1):

$$\hat{\psi} = \log \frac{r_1/(n_1 - r_1)}{r_0/(n_0 - r_0)} \tag{2}$$

Adjusted (pooled) estimator based on *strata* $s \in S$ defined by background variable(s) W:

$$\hat{\psi} = \frac{\sum_s \hat{\psi}_s/v_s}{\sum_s 1/v_s}. \tag{3}$$

$\hat{\psi}_s$ and v_s are the point and asymptotic variance estimates (using (2)) in stratum s.
The variance estimate of $\hat{\psi}$ is $(\sum_s 1/v_s)^{-1}$.

Consistency ($\hat{\psi}_s = \hat{\psi}$ for all s) can be tested by a χ^2 test statistic $\sum_s (\hat{\psi}_s - \hat{\psi})^2/v_s$.

Different models

Population model Joint distribution $\mathbb{P}\{Y, X\}$ of outcome Y and exposure X in the population:

	$Y=0$	$Y=1$
$X=0$	π_{00}	π_{01}
$X=1$	π_{10}	π_{11}

Formal interpretative (or inverse) model Prospective: Outcome given the exposure:

	$Y=0$	$Y=1$	$\mathbb{P}\{Y=y \mid X=x\}$
$X=0$	π_{00}	π_{01}	$\pi_{01}/(\pi_{01}+\pi_{00})$
$X=1$	π_{10}	π_{11}	$\pi_{11}/(\pi_{11}+\pi_{10})$

Sampling model Retrospective: Exposure given the outcome:

	$Y=0$	$Y=1$
$X=0$	π_{00}	π_{01}
$X=1$	π_{10}	π_{11}
$\mathbb{P}\{X \mid Y\}$	$\pi_{10}/(\pi_{10}+\pi_{00}) =: \theta_0$	$\pi_{11}/(\pi_{11}+\pi_{01}) =: \theta_1$

Note that $\mathrm{OR} = \pi_{11}\pi_{00}/(\pi_{01}\pi_{10}) = \theta_1/(1-\theta_1) \,/\, (\theta_0/(1-\theta_0))$.

Variance estimator of log OR

Consider number of cases n_1 fixed, and conditional probability $\theta_1 = \mathbb{P}\{X=1 \mid Y=1\}$. Then number of exposed $R_1 \sim \mathrm{Binomial}(n_1, \theta_1)$.
However, variance estimation for log odds $\log(R_1/(N_1 - R_1))$ is simpler by considering Poisson distributed random variables V_i (mean γ_i, $i \in \{0, 1\}$).

1. Binomial distribution
   $[V_1 \mid V_0 + V_1 = v] \sim \mathrm{Binomial}(v, \gamma_1/(\gamma_0 + \gamma_1))$.
2. Asymptotical normality Asymptotically ($\gamma_i \to \infty$)
   $(\log V_i - \log \gamma_i)/\sqrt{1/V_i} \to \mathrm{N}(0, 1)$ (delta method for $\log(V_1/\gamma_1)$).
   Linear combinations $c_0 \log V_0 + c_1 \log V_1$ (and $d_0 \log V_0 + d_1 \log V_1$) is normally distributed with mean $\sum_i c_i \log V_i$, variance $\sum_i c_i^2/\gamma_i$ and covariance $\sum_i c_i d_i/\gamma_i$.
3. Uncorrelated contrast If $\sum_i c_i = 0$ (a contrast) then $\mathrm{Cov}(\sum_i c_i \log V_i, \sum_i V_i) = 0$.

Variance estimator of log OR ...

Assume R_1 and $N_1 - R_1$ independent Poisson r.v.'s:

Results 1 and 2 imply asymptotic mean $\theta_1/(1-\theta_1)$ and variance $1/(n_1\theta_1) + 1/\{n_1(1-\theta_1)\}$.

Result 3 implies that asymptotical variance is the same conditionally or unconditionally n_1.

As the same calculations apply also for controls, we get asymptotic variance estimate for log OR:

$$\frac{1}{r_1} + \frac{1}{n_0 - r_0} + \frac{1}{n_1 - r_1} + \frac{1}{r_0} \tag{4}$$

Logistic regression models for more general adjusted CC analyses

It can be shown that a standard logistic regression model can be applied in CC analysis assuming prospective design i.e. using Y as the outcome. Exposure X, and confounders and background variables W can be included in the model as covariates.

- The intercept term based on CC data does not equal to the intercept based on prospective analysis based on data representing the population.
- Relationships of variables W do not necessarily represent those in the population. E.g. if risk factors X_1 and X_2 are independent in the population, but affect outcome Y, then CC sampling based on $Y = 1$ creates association between X_1 and X_2 in the CC data (recall also collider nodes in the work of Pearl).
- Stratified analyses can also be conducted prospectively.

Generally regression models do not have the property of analyzing retrospective data as a prospective data.

Matched case-control study

Binary exposure and binary outcome

For each case one (or several) controls are selected individually.
Matching should be based on variables W which are causally prior to the exposure X.
There can be four possible pairs of exposure.
For each pair u the likelihood of $X_{u;0} = x_{u;0}$ and $X_{u;1} = x_{u;1}$ is (assuming logistic regression model with parameters β_u and $\beta_u + \beta_0$, respectively):

	Case $x_{u;1}$	Control $x_{u;0}$	Likelihood term
Concordant	0	0	K
Discordant	1	0	$\exp\{\beta_u + \beta_0\}K$
Discordant	0	1	$\exp\{\beta_u\}K$
Concordant	1	1	$\exp\{\beta_u + \beta_0\}\exp\{\beta_u\}K$

where

$$K := \frac{1}{(1+\exp\{\beta_u + \beta_0\})(1+\exp\{\beta_u\})}.$$

β_0 is the parameter of interest, and β_u are nuisance parameters.

Matched case-control designs

The likelihood terms are

$$\frac{\exp\{\beta_u x_{u;0}\}}{1+\exp\{\beta_u\}} \frac{\exp\{(\beta_u + \beta_0)x_{u;1}\}}{1+\exp\{\beta_u + \beta_0\}} \tag{5}$$

Note that $x_{u;0} + x_{u;1} =: x_{u;\cdot}$ is the minimal sufficient statistic for β_u.
Conditioning on $x_{u;\cdot}$, the likelihood does not depend on β_u.
The concordant pairs (0,0) and (1,1) have conditional probability equal 1:

$$\mathbb{P}\{X_{u;0} = 0 \,|\, X_{u;0} + X_{u;1} = 0\} = 1$$
$$\mathbb{P}\{X_{u;1} = 1 \,|\, X_{u;0} + X_{u;1} = 2\} = 1$$

thus these pair do not contain information about β_0.
Only discordant pairs contain information about β_0:

$$\mathbb{P}\{X_{u;1} = x_{u;1} \,|\, X_{u;0} + X_{u;1} = 1\} = \frac{\exp\{\beta_u + \beta_0 x_{u;1}\}K}{\exp\{\beta_u\}K + \exp\{\beta_u + \beta_0\}K} \tag{6}$$

Matched case-control designs

The conditional odds ratio based on (6) (after applying definition of odds $O := p/(1-p)$) is $\exp\{\beta_0\}$.
For $n_D := n_{10} + n_{01}$ observed discordant pairs, where n_{10} is the number of pairs with exposed cases and unexposed controls, the conditional likelihood equals the binomial likelihood

$$\left(\frac{\exp\{\beta_0\}}{1+\exp\{\beta_0\}}\right)^{n_{10}} \left(\frac{1}{1+\exp\{\beta_0\}}\right)^{n_{01}} \tag{7}$$

Point and large sample variance estimates for $\beta_0 = \log \mathrm{OR}$ are

$$\hat{\psi} = \log \frac{n_{10}}{n_{01}} \quad \text{and} \quad \frac{1}{n_{10}} + \frac{1}{n_{01}}.$$

Conditional logistic regression model can be applied to adjust for background variables W.