Case-Cohort Design and Secondary Analysis

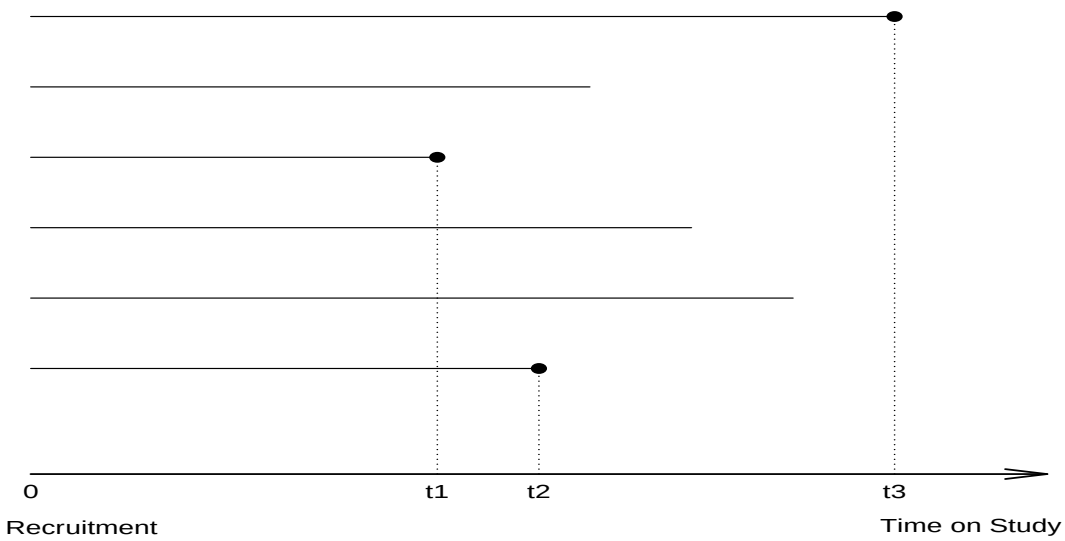
Yujie Zhong and Richard J. Cook

Department of Statistics and Actuarial Science, University of Waterloo

Introduction

Why?

Cohort studies for involving chronic disease aim to examine the effect of risk factors on incidence of disease. They typically require a extended follow-up of a large cohort of individuals to ensure a certain precision and power to detect the risk factors effect.



Controls: Follow-up ends without •

Cases: Follow-up ends with •

Why?

The standard cohort design is

- Expensive: Requires assembly of all covariate (exposure) histories
- Inefficient: Measuring exposure in many controls is wasteful
- Time-consuming: Requires extended follow-up to observe the development of the condition of interest

Why?

Example

Multiple risk factor intervention trial [1]
This trial is to determine whether a special intervention program would result in a significant reduction in mortality from coronary heart disease.

- randomized 12,866 subjects, and followed up for seven years
- cost more than $100 million
- only 2% of MRFIT men experienced the primary endpoint of coronary heart disease mortality.

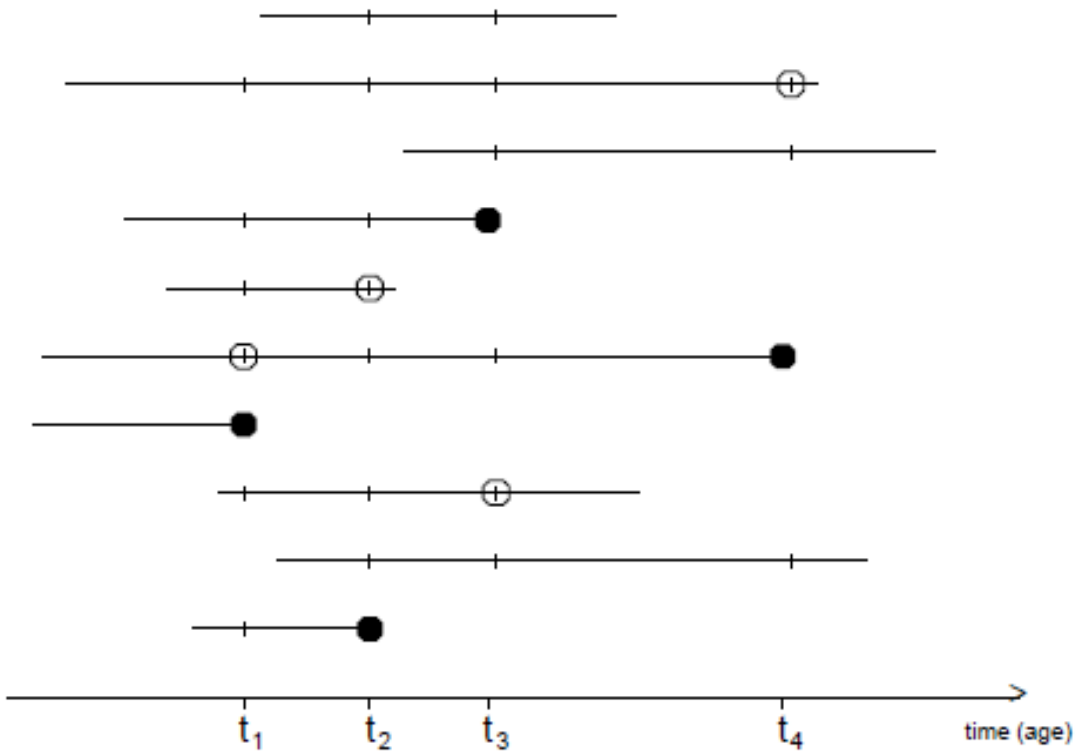
Alternative more cost-effective designs are needed:

- Nested case-control design (Thomas, 1976)
- Case-cohort design (Prentice, 1986)

∗ MRFIT Research Group (1982)

Nested Case-Control Design

In a nested case-control design, a number of "controls" from those at risk at the failure time of each case is selected.



*Cases: • Controls: ∘

∗ Samuelsen(2005)

Case Cohort Design

In the case-cohort design, a random sample (subcohort) is chosen from the cohort and covariate/exposure data is collected for

- this subcohort
- all "cases" outside subcohort

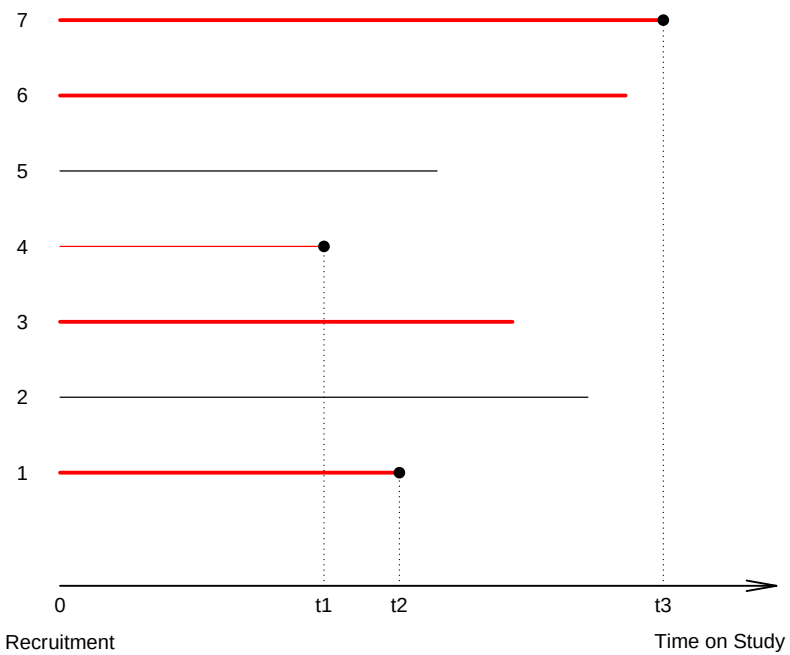
Subcohort: Red lines

$$\mathcal{C} = \{1, 2, \ldots, 7\}$$
$$\mathcal{S} = \{1, 3, 6, 7\}$$

Case Cohort Design

In the case-cohort design, a random sample (subcohort) is chosen from the cohort and covariate/exposure data is collected for

- this subcohort
- all "cases" outside subcohort



Subcohort: Red lines

Cases: •

$$\mathcal{C} = \{1, 2, \ldots, 7\}$$
$$\mathcal{S} = \{1, 3, 6, 7\}$$
$$\mathcal{S} \cup \mathcal{D} = \{1, 3, 4, 6, 7\}$$

Advantages of Case Cohort Design

- Much lower total cost for measurement of exposure
- More efficient than analysis based on subcohort only
- Data from the subcohort selected in advance of follow-up can be useful to study time varying markers
- Enables analysis for a range of disease endpoints...

Notation

- T_i - event time
- C_i - censoring time
- $X_i = \min(T_i, C_i)$
- $\delta_i = \mathrm{I}(T_i \leq C_i)$
- $Y_i(t) = \mathrm{I}(T_i \geq t)$
- $\bar{Y}_i(t) = \mathrm{I}(t \leq C_i) Y_i(t)$
- $N_i(t)$ is a counting process
- $Z_i(u)$ is $p \times 1$ covariate vector

Full Cohort: $\mathcal{C} = \{1, \ldots, N\}$

Subcohort: $\mathcal{S} = \{i_1, \ldots, i_n;\ 1 \leq i_j \leq N\}$ of size n with $\bar{\mathcal{S}} = \mathcal{C} \backslash \mathcal{S}$

Cases: $\mathcal{D}(t) = \{i \in \mathcal{C} :\ N_i(t) \neq N_i(t^-)\}$

Recall: Estimation via Partial likelihood (Cox, 1972)

Under the Cox model

$$\lambda(t; Z(u), 0 \leq u \leq t) \quad = \quad \lambda_0(t) \exp(Z^{'}(t)\beta)$$

The partial likelihood and score equation are

$$L(\beta) \propto \prod_{i=1}^{N} \left[\frac{\bar{Y}_i(t_i) \exp(Z_i^{'}(t_i)\beta)}{\sum_{j \in \mathcal{C}} \bar{Y}_j(t_i) \exp(Z_j^{'}(t_i)\beta)} \right]^{\delta_i}$$

$$U(\beta) = \sum_{i=1}^{N} \int_0^{\infty} \bar{Y}_i(s) \left[Z_i(s) - \frac{\sum_{j \in \mathcal{C}} \bar{Y}_j(s) \exp(Z_j^{'}(s)\beta) Z_j(s)}{\sum_{j \in \mathcal{C}} \bar{Y}_j(s) \exp(Z_j^{'}(s)\beta)} \right] dN_i(s)$$

Estimation via Pseudo-likelihood (Prentice, 1986)

The pseudo-likelihood and score equation for a case-cohort study are

$$\tilde{L}(\beta) \propto \prod_{i=1}^{N} \left[\frac{\bar{Y}_i(t_i) \exp(Z_i'(t_i)\beta)}{\sum_{j \in \mathcal{S} \cup D(t_i)} \bar{Y}_j(t_i) \exp(Z_j'(t_i)\beta)} \right]^{\delta_i} \quad (1)$$

$$\tilde{U}(\beta) = \sum_{i=1}^{N} \int_0^{\infty} \bar{Y}_i(s) \left[Z_i(s) - \frac{\sum_{j \in \mathcal{S} \cup D(s)} \bar{Y}_j(s) \exp(Z_j'(s)\beta) Z_j(s)}{\sum_{j \in \mathcal{S} \cup D(s)} \bar{Y}_j(s) \exp(Z_j'(s)\beta)} \right] dN_i(s) \quad (2)$$

Estimation via Pseudo-likelihood (Prentice, 1986)

The pseudo-likelihood estimator $\tilde{\beta}$ solves $\tilde{U}(\tilde{\beta}) = 0$, and has following asymptotic distribution

$$\sqrt{N}(\tilde{\beta} - \beta) \longrightarrow N(0, \Sigma^{-1}(\Sigma + \Delta)\Sigma^{-1})$$

- Σ is consistently estimated by information matrix generated by pseudo-likelihood
- Δ reflects the contribution of the covariance among score components induced by the case-cohort sampling scheme

Evolution of Research on The Case-Cohort Design

Recent Methodologic Advances

- Various estimating functions have been proposed to give different estimators
- Approaches to variance estimation have been explored
- Alternative sampling schemes

A Comparison of Different Estimation Methods

A Comparison of Different Estimation Methods

[I] Weighting Methods

Recall the pseudo-likelihood

$$\tilde{L}(\beta) \propto \prod_{i=1}^{N} \left[\frac{\bar{Y}_i(t_i) \exp(Z_i'(t_i)\beta)}{\sum_{j \in \mathcal{S} \cup D(t_i)} \bar{Y}_j(t_i) \exp(Z_j'(t_i)\beta)} \right]^{\delta_i}$$

This can be modified by introducing weights to reflect the sampling scheme

$$\tilde{L}_w(\beta) \propto \prod_{i=1}^{N} \left[\frac{\bar{Y}_i(t_i) \exp(Z_i'(t_i)\beta)}{\sum_{j \in \mathcal{C}} w_j(t_i) \bar{Y}_j(t_i) \exp(Z_j'(t_i)\beta)} \right]^{\delta_i}$$

where $w_j(t_i)$ is weight for subject j at time t_i.

A Comparison of Different Estimation Methods

[I] Weighting Methods

Carefully examine the denominator of $\tilde{L}_w(\beta)$

$$\bar{Y}_i(t_i)w_i(t_i)e^{Z_i'(t_i)\beta} + \sum_{j\in\mathcal{S}\setminus\{i\}} \bar{Y}_j(t_i)w_j(t_i)e^{Z_j'(t_i)\beta} + \sum_{j\in\bar{\mathcal{S}}\setminus\{i\}} \bar{Y}_j(t_i)w_j(t_i)e^{Z_j'(t_i)\beta} \quad (3)$$

Table1: Weights for Different Methods

Individual j		Weight at time t_i		
Status	Group	Prentice[1]	SP[2]	Barlow[3]
Control	$\bar{\mathcal{S}}$	0	0	0
Control	$\mathcal{S}$	1	1	$1/P(j \in \mathcal{S})$
Case	$\bar{\mathcal{S}}$	1	0	$\mathrm{I}(\mathrm{t_i} = \mathrm{t_j})$
Case	$\mathcal{S}$	1	1	$\mathrm{I}(\mathrm{t_i} = \mathrm{t_j}) + \mathrm{I}(\mathrm{t_i} \neq \mathrm{t_j})/\mathrm{P}(\mathrm{j} \in \mathcal{S})$

[1] Prentice (1986); [2] Self and Prentice (1988); [3] Barlow [1994]

Unbiasness of Weighted Score Equations

- $R_j(t) = \mathrm{I}(\text{subject j contributes at t})$
- $\pi_j(t) = \bar{Y}_j(t)dN_j(t) + \bar{Y}_j(t)(1 - dN_j(t))P(j \in \mathcal{S}|\bar{Y}_j(t) = 1, Z_j(t))$
- $w_j(t) = R_j(t)/\pi_j(t)$

Then score equations for case-cohort design are

$$U_1(d\Lambda_0(t), \beta) = \sum_{j=1}^{N} \left[\frac{R_j(t)}{\pi_j(t)}\right] \left\{\bar{Y}_j(t)\left[dN_j(t) - d\Lambda_j(t|Z_j(t))\right]\right\} \tag{4}$$

$$U_2(d\Lambda_0(t), \beta) = \sum_{j=1}^{N} \int_0^{\infty} \left[\frac{R_j(s)}{\pi_j(s)}\right] \left\{\bar{Y}_j(s)Z_j(s)\left[dN_j(s) - d\Lambda_0(s)e^{Z_j'(s)\beta}\right]\right\} \tag{5}$$

Unbiasness of Weighted Score Equations

Conditional expectation of $R_j(s)$ given $\{\bar{Y}_j(s) = 1, dN_j(s), Z_j(s)\}$,

$$
\begin{aligned}
& E[R_j(s)|\bar{Y}_j(s) = 1, dN_j(s), Z_j(s)] \\
= \; & \bar{Y}_j(s)(1 - dN_j(s))P(R_j(s) = 1|\bar{Y}_j(s) = 1, dN_j(s) = 0, Z_j(s)) \\
& + \bar{Y}_j(s)dN_j(s)P(R_j(s) = 1|\bar{Y}_j(s) = 1, dN_j(s) = 1, Z_j(s)) \\
= \; & \bar{Y}_j(s)(1 - dN_j(s))P(j \in \mathcal{S}|\bar{Y}_j(s) = 1, dN_j(s) = 0, Z_j(s)) \\
& + \bar{Y}_j(s)dN_j(s) \\
= \; & \pi_j(s)
\end{aligned}
$$

Unbiasness of Weighted Score Equations

$$
\begin{aligned}
E[U_1(d\Lambda_0(t),\beta)] &= E[E\{U_1(d\Lambda_0(t),\beta)|\bar{Y}_j(t)=1, dN_j(t), Z_j(t)\}] \\
&= E\left[\sum_{j=1}^{N} \bar{Y}_j(t)\left[dN_j(t) - d\Lambda_j(t|Z_j(t))\right]\right] \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
E[U_2(d\Lambda_0(t),\beta)] &= E[E\{U_2(d\Lambda_0(t),\beta)|\bar{Y}_j(t)=1, dN_j(t), Z_j(t)\}] \\
&= E\left\{\sum_{j=1}^{N}\int_0^{\infty} \bar{Y}_j(s)Z_j(s)\left[dN_j(s) - d\Lambda_0(s)e^{Z_j'(s)\beta}\right]\right\} \\
&= 0
\end{aligned}
$$

Estimation Methods

[II] Weighted Estimating Equations: Chen and Lo (1999) suggested a class of estimating functions that in many cases improved efficiency.

[III] Cohort study for missing data: Lin and Ying (1993) proposed an approximate partial likelihood method for missing covariate problem under Cox model, and applied that in a case-cohort study.

Variance Estimation

Variance Estimation

Self and Prentice (1988)

- Derive the asymptotic variance

Wacholder et al (1989)

- Bootstrap variance estimation

Barlow (1994)

- Robust variance estimation

Sampling Schemes

Sampling Scheme

Borgan et al. (2000) proposed exposure stratified case-cohort design and three estimators are suggested.

- Stratification is based on easily observable inexpensive covariates available for the entire cohort $\mathcal{C}$
- Stratified sampling design analyzed using the weighted version of the pseudo-likelihood estimating method can dramatically increase statistical efficiency.
- Software: R/S-PLUS and SAS (Samuelsen et al. 2007)

```
cch(formula, data = sys.parent(), subcoh, id, stratum=NULL, cohort.size,
    method =c("Prentice","SelfPrentice","LinYing","I.Borgan","II.Borgan"),
    robust=FALSE)
```

Available Software

- R/S-PLUS
  - 'cch' function in 'survival' package in *R* fit Cox model for case-cohort data based on three methods: Prentice (1986), Self and Prentice (1988) and Lin and Ying (1993) and give asymptotic variance or robust variance.
  - For stratified case-cohort data, 'cch' can only use Borgan I and Borgan II. Although Borgan III method is not available in 'survival' package, Cologne et al. (2012) share S-PLUS code for this method.
- SAS

  'PHREG' procedure gives parameter estimates, and 'IML' can be used to compute the robust variance estimate; see Langholz and Jiao (2007)

A Simulation Study

- Aim: Examine the relative efficiency and cost savings of case-cohort vs cohort design.
- Setting
  - Cohort size: $N = 5000$
  - Subcohort size: $n = 500,\ 1000,\ 2500$
  - Simulation times: $m = 1000$
  - Follow-up interval: $[0, 1]$
  - Covariate/Exposure: $Z \sim N(0, 1)$
  - Disease onset time: $T|Z \sim \exp(\lambda e^{Z\beta})$
    - –– Hazard ratio: $\exp(\beta) = 1.25$
    - –– Incidence rate: 10%, 40%
- Analyses
  - Full Cohort Analysis
  - Case-Cohort I (Prentice, 1986)
  - Case-Cohort II (Lin and Ying, 1993)

A Simulation Study

Table 2: Estimation results (10% Disease Incidence, $N = 5000$)

	n=500			n=1000			n=2500		
Method	BIAS	ASE	ESE	BIAS	ASE	ESE	BIAS	ASE	ESE
Cohort	1e-05	.044	.045	.002	.044	.046	.003	.044	.044
CCP	.002	.069	.062	.003	.057	.056	.003	.047	.047
CCLY	.002	.063	.062	.003	.053	.055	.003	.047	.046

Table 3: Estimation results (40% Disease Incidence, $N = 5000$)

	n=500			n=1000			n=2500		
Method	BIAS	ASE	ESE	BIAS	ASE	ESE	BIAS	ASE	ESE
Cohort	-15e-04	.023	.024	-4e-05	.023	.023	4e-05	.022	.023
CCP	-6e-04	.021	.054	3e-04	.043	.040	-1e-04	.029	.028
CCLY	-9e-04	.048	.050	-3e-04	.036	.037	-4e-05	.027	.027

A Simulation Study: Consider Cost Saving

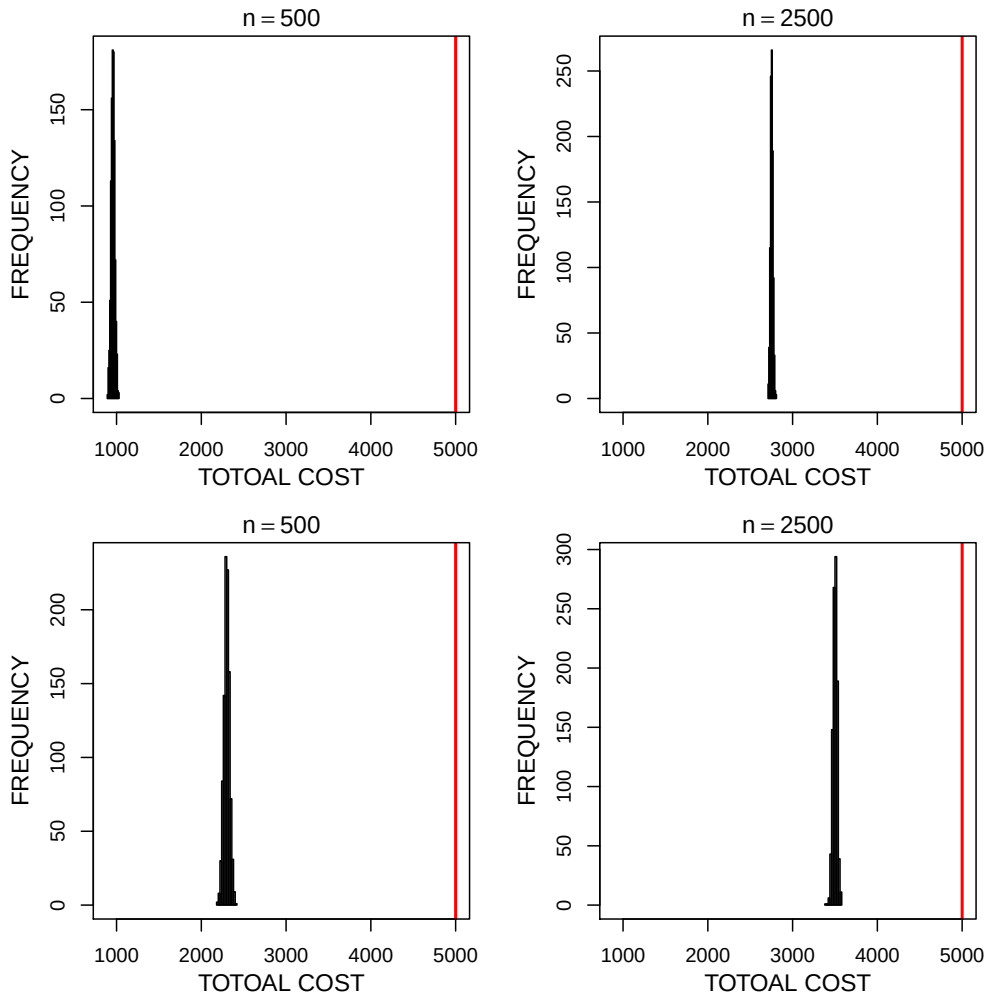


Fig1: Histogram for cost for case-cohort design (upper: 10%, bottom: 40%)

Multiple Disease Outcomes

Able to study multiple outcomes individually without having to obtain separate sets of comparison subjects.
However, this aspect is seldom illustrated or examined in literature.

- Sørensen and Andersen (2000)
  - investigate competing risk of multiple outcomes for case-cohort design
  - when all case groups are compared with the same sub-cohort, a correlation is induced between estimated exposures effects on different outcomes.
  - Correlations increase with smaller subcohort sampling fractions.
- Kang and Cai (2009)
  - Simultaneously model the times to different events to compare the effect of a risk factor on different types of diseases
  - Marginal proportional hazards regression models are proposed for case-cohort studies with multiple disease outcomes.

Secondary Analysis

Secondary Analysis

- Uses existing data to investigate research questions other than the main ones for which the data were originally gathered (Hulley et al., 2007)
- Can supplement the primary analysis and guide scientific inquiry, or it might be the primary interest for subsequent research

There has been little discussion of statistical issues arising in secondary analysis of case-cohort data

Secondary Analysis for Case-Control Design

Common methods:

[1] Use only the controls

[2] Use only the cases

[3] Ignore sampling scheme, use both cases and controls

[4] Joint analysis of cases and controls adjusted for disease status

None of the above analysis methods is statistically correct!

- Cases and controls are selected at different rates and do not represent the population ([3])
- Association between risk factors and secondary trait in the sampled subgroups can be different from that in the general population ([1],[2],[4])

Secondary Analysis for Case-Control Design

Lin and Zeng (2009)

- Proposed a new likelihood based method that reflects the case-control sampling (retrospective form and based on disease status)
- After specifying $P(Y_i|X_i)$ (linear or logistic regression) and $P(D_i = 1|X_i, Y_i)$ (logistic regression), the likelihood is

$$L = \prod_{i=1}^{n} P(Y_i, X_i|D_i)$$

where Y_i is secondary trait, D_i is case-control status, and X_i are covariate.

$*$ Lee, McMurchy and Scott (1997); Jiang, Scott and Wild (2006)

Secondary Analysis For Case-Cohort Design

Copula Function

- (T_1, T_2): two correlated event times with survivor functions $\mathcal{F}_1(t)$ and $\mathcal{F}_2(t)$, respectively.

We specify a joint model via a copula function

- A copula function in 2 dimensions is a multivariate distribution on the unit $[0,1]^2$ whose margins are all uniform over $[0,1]$:

$$H(u_1, u_2; \phi) \quad = \quad P(U_1 \leq u_1, U_2 \leq u_2)$$

- Joint survivor function for (T_1, T_2) derived from copula is

$$\mathcal{F}(t_1, t_2) \quad = \quad H(\mathcal{F}_1(t_1), \mathcal{F}_2(t_2); \phi)$$

Secondary Analysis for Case-cohort Design

- $\bar{Y}_{i2}(t) = I(T_{i2} \geq t,\ T_{i2} \geq C_i)$
- $\mathcal{D}_1(t) = \{i;\ dN_1(t) \neq dN_1(t^-)\}$
- $Z_{i2}(t)$: covariate vector

The key here is how to revise the second term of score equations:

$$\frac{\sum_{i\in\mathcal{S}} \bar{Y}_{i2}(t)Z_{i2}(t)e^{\beta_2' Z_{i2}} + \sum_{i\in\bar{\mathcal{S}}} \bar{Y}_{i2}(t)Z_{i2}(t)e^{\beta_2' Z_{i2}(t)}}{\sum_{i\in\mathcal{S}} \bar{Y}_{i2}(t)e^{\beta_2' Z_{i2}(t)} + \sum_{i\in\bar{\mathcal{S}}} \bar{Y}_{i2}(t)e^{\beta_2' Z_{i2}(t)}}$$

$$\Downarrow$$

$$\frac{\sum_{i\in\mathcal{S}} w_i(t)\bar{Y}_{i2}(t)Z_{i2}(t)e^{\beta_2' Z_{i2}(t)} + \sum_{i\in\mathcal{D}_1(t)\cap\bar{\mathcal{S}}} w_i(t)\bar{Y}_{i2}(t)Z_{i2}(t)e^{\beta_2' Z_{i2}(t)}}{\sum_{i\in\mathcal{S}} w_i(t)\bar{Y}_{i2}(t)e^{\beta_2' Z_{i2}(t)} + \sum_{i\in\mathcal{D}_1(t)\cap\bar{\mathcal{S}}} w_i(t)\bar{Y}_{i2}(t)e^{\beta_2' Z_{i2}(t)}}$$

What $w_i(t)$ should we use here?

Secondary Analysis for Case-Cohort Design

Necessities for weights $w_i(t)$:

- reflecting the sampling scheme
- making the estimating equations (asymptotically) unbiased

Option 1: $w_i(t) = \mathrm{I}(\mathrm{i} \in \mathcal{S})$

- only use information on subcohort
- estimating equation is unbiased

Secondary Analysis for Case-Cohort Design

Our aim: make use of available information outside the subcohort.

- For $i \in \mathcal{S}$, $w_i(t)$ can be the same as before, i.e. 1 or inverse probability weights
- The key question is how to choose $w_i(t)$ for $i \in \mathcal{D}_1(t) \cap \bar{\mathcal{S}}$

*Option 2:

$$w_i(t) = \mathrm{I}(T_{i1} \leq C_i)/P(T_{i1} \leq C_i | T_{i2} \geq t), \qquad \text{for } i \in \mathcal{D}_1(t) \cap \bar{\mathcal{S}}$$

This can be estimated parametrically or nonparametrically.

Secondary Analysis for Case-Cohort Design

Next Step:

- Prove the (asymptotic) unbiasness of the implied estimating equation
- Find the asymptotic distribution of estimator
- Find the estimates of parameters and asymptotic variance
- Simulation and application

Summary & Comments

- Case-cohort designs offer a cost-effective method of estimating the effects of risk factors in setting with low disease incidence and covariates which are costly to measure
- Since all cases are used there is little loss in precision or power
- Potential use in genomic and other biomarker studies is considerable
- A variety of weighting schemes can be explored to extract as much information from data as possible
- Methods for conducting valid secondary analyses are on going