

Epidemiologically valid and statistically efficient designs for biobank-based studies

Esa Läärä

Department of Mathematical Sciences,
University of Oulu, Finland

and

Finnish Cancer Registry, Helsinki

Nordic Research Seminar on Biobanking
Malmö, 28 August, 2006

Outline

Validity & efficiency

Cohort design

Case-control designs

Nested CC

Case-cohort

Estimation and precision

Matching

Conclusion

References

Borgan Ø, Samuelssen S-O. A review of cohort sampling designs for Cox's regression model: Potentials for epidemiology. *Norsk Epidemiologi* 2003; 13(2):239-248.

dos Santos Silva, I. *Cancer Epidemiology: Principles and Methods*. International Agency for Research on Cancer, Lyon 1999.

Rundle AG, Vineis P, Ahsan, H. Design Options for Molecular Epidemiology Research within Cohort Studies. *Cancer Epidemiol Biomarkers Prev* 2005; 14(8): 1899-1907.

Epidemiologic parameters & estimation

Epidemiologic study is a *measurement exercise*.

Object: some **parameter** of interest, like

- hazard rate ratio (HR, "relative risk") of cervix cancer between smokers and non-smokers,

Result: **Estimate** of the parameter, computed from data (e.g. incidence rate ratio IRR, or exposure odds ratio EOR)

Estimation of parameter is prone to error:

$$\begin{aligned} \text{estimate} &= \text{true parameter value} \\ &+ \text{systematic error (bias)} \\ &+ \text{random error} \end{aligned}$$

Example: Smoking and cervix cancer (Simen Kapeu 2006)

Study population, follow-up, design, and measurements

- Joint cohort of $N \approx 550\,000$ from 3 biobanks,
- 1st entries & blood samples in 1974, follow-up till 1994.
- 171 cervix cancer patients, 496 controls.
- Each case matched with 3 controls by age (± 2 y), region, and storage time of blood sample.
- Blood samples analyzed for cotinine, antibodies to HPV16, 18, 33, and *C. trachomatis*.

Main result: Adjusted OR = 1.6 (95% CI: 1.0 to 2.4) for high cotinine levels > 242.6 ng/ml vs. low level < 3.0 ng/ml.

What does all this mean?

Bias and random error

Sources of bias

- confounding – non-comparability of exposure groups,
- measurement error & misclassification,
- non-response, loss to follow-up, incomplete data,
- sampling, selection, . . .

Sources of random error

- biological variation between and within individuals,
- measurement variation,
- sampling (random or non-random),
- allocation of exposure (randomized or non-randomized).

Validity – unbiasedness

An epidemiologic study is **valid**, when its

- design and methods will provide an
- *unbiased* estimate of the
- parameter (like HR) of interest.

Unbiased estimation

- ⇔ the estimate (EOR) would equal the true parameter value (HR), if the study had perfect *precision* (infinite size, no random error)

Example: If the true HR for high vs. low cotinine level were 2.0, this value would be exactly obtained by our estimate EOR, if we had unlimited amount of data, and if our design were valid.

(Of course, by exceptional luck we could get an EOR of 2.0 also with typical amount of data even with biased design!)

Outline of a cohort study

- Cohort is defined: eligibility criteria of membership.
- Risk factors of interest, confounders & modifiers are measured in all cohort members.
- New incident cases of outcome (e.g. cancer) are identified during *follow-up* from *entry* time till *exit* time.
- Incidence rates = cases/person-time in exposure groups, and incidence rate ratios (IRR) between them are computed
- Confounding and modification are controlled by
 - stratification & Mantel-Haenszel methods,
 - regression modelling: Poisson, or proportional hazards model (Cox 1972)

Crude estimation of hazards and their ratios

Simplified summarization of results in a cohort study:

	exposed	unexposed
cases	D_1	D_0
person-time	Y_1	Y_0
incidence rate	$R_1 = D_1/Y_1$	$R_0 = D_0/Y_0$

The hazard rate ratio (HR) is estimated as the ratio of empirical incidence rates

$$IRR = \frac{R_1}{R_0} = \frac{D_1/Y_1}{D_0/Y_0} = \frac{D_1/D_0}{Y_1/Y_0}$$

Precision and efficiency

Precision of an estimate: lack of random error

Random error is measured by

- variance or *standard error* (SE) of the estimate, or
- *confidence interval* (CI) of the parameter

Efficiency of a design:

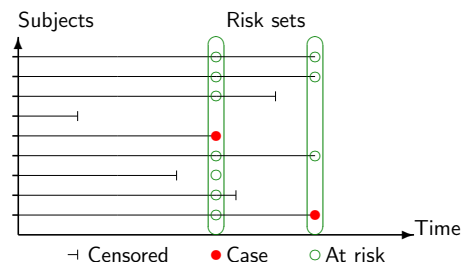
- ⇔ Ability to provide a precise estimate with given data.

Design A is more efficient than design B, if

- with same amount of data the estimate from A has a smaller random error than that from B, or
- smaller amount of data is needed by design A to obtain the same precision than obtained by B.

Cohort follow-up and risk sets

Each member of the cohort provides exposure data for all cases, as long as the cohort member is **at risk**, *i.e.* alive, not censored & free from outcome.



Precision and efficiency of HR estimation

Precision depends on the numbers of cases.

The variance of the logarithm of IRR

$$= \frac{1}{D_1} + \frac{1}{D_0} = \frac{1}{\text{no. exposed cases}} + \frac{1}{\text{no. unexposed cases}}$$

- Provides the basis for calculating an approximate CI for HR.
- Does not depend on group sizes (or person-times) as such, even though these were millions.
- Yet, for rare diseases with low rates, large cohorts are needed to obtain sufficiently many cases.

Problems with full sampling of cohort data

Collection & processing of exposure and covariate data

- slow and expensive in large cohorts
- not feasible for certain data, e.g.
 - biological measurements,
 - dietary diaries,
 - occupational exposure histories.

Question:

- *Can we obtain equally valid estimates of HRs with nearly as good precision by some other strategies?*

Case-control designs

General principle:

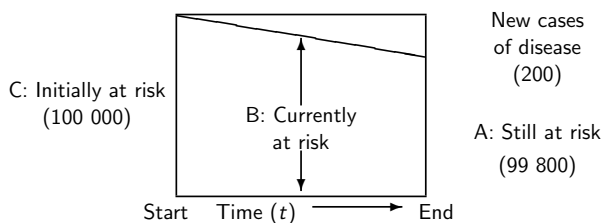
Selection of study subjects from a given *study population* is *stratified by the outcome* (disease) under study.

Study Population

- *Ideally*: subjects who would be included as cases, if they had got the outcome during the study
= *Source population* of the cases.
- In *hospital-based* case-control studies often difficult to identify.
- No problem in well-defined cohorts that are under practically complete follow-up for outcome diseases and total mortality.

Follow-up of a cohort & sampling frames for controls

Simplified ideal situation: Complete follow-up of an initially healthy subjects at risk with no losses or censorings during the study time.



Estimation of HR (cont'd)

The incidence rate ratio in a cohort-study can be expressed:

$$\text{IRR} = \frac{D_1/D_0}{Y_1/Y_0} = \frac{\text{cases: exposed / unexposed}}{\text{person-times: exposed / unexposed}}$$

$$= \frac{\text{exposure odds in cases}}{\text{exposure odds in p-times}} = \text{exposure odds ratio}$$

= Exposure distribution in cases vs. that in the whole cohort !

Implication for more efficient design:

- Collect exposure data on all cases, but
- Estimate person-time distribution by sampling "control" subjects, on whom exposure data will be collected, from the cohort members at risk.

Obtaining exposure data

Data on risk factors are collected separately from

- (I) **Case group**: All (or high % of) subjects getting the outcome in the study population during the period.
- (II) **Control group**:
 - A *sample* (simple or stratified) of subjects from the study population, 1 to 4 (rarely more) per case.
 - Controls must be *at risk* (alive, under follow-up & free from outcome) at a defined time point,
 - different *sampling schemes* or *designs* available.

Sampling designs for controls

A: Traditional, "case-noncase" design

- Controls chosen from those cohort members still at risk *at the end* of the follow-up.
- Requires complete follow-up (no censoring) over the same fixed risk period for all subjects.

B: Nested case-control design (NCC)

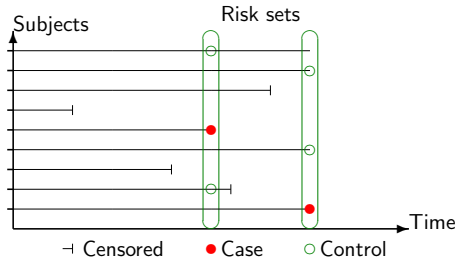
- Also called *density sampling*, *time-matched s.* or *concurrent s.*
- For each case, 1 or more controls drawn from those cohort members at risk *at the time of diagnosis* of the case,

C: Case-cohort design (CC)

- The control group, also called *subcohort*, is a random sample of the whole cohort *at the start* of the follow-up.

Nested case-control design

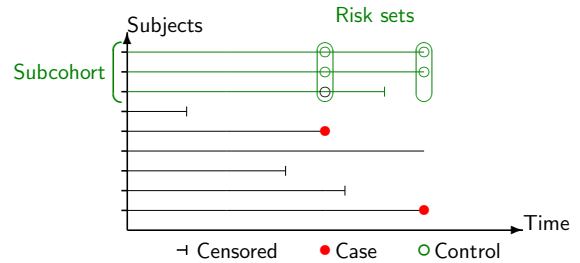
When a new case occurs, a set of controls (here 2/case) are sampled from its *risk set* = those at risk at the diagnosis time.



NB. A control for a previous case can later on become a case, too.

Case-cohort design

Subcohort: Sample of the whole cohort randomly selected at the outset. Serves as control group for all cases.



NB. A subcohort member can become a case, too.

Use of different designs

Design A: Traditional

- Common in acute diseases & perinatal epidemiology,
- Not suitable for cancer studies with variable follow-up times due to staggered entry and extensive censoring.

Design B: Nested case-control

- Most popular design in chronic disease epidemiology,

Design C: Case-cohort

- Good alternative to NCC in certain circumstances,

Designs B and C allow statistically valid estimation of the hazard ratio HR without any "rare disease" assumption.

Estimation of "relative risk"

Simplified summary of results of a case-control study

	exposed	unexposed	total
cases	D_1	D_0	D
controls	C_1	C_0	C

From these data, exposure odds ratio (EOR) between cases and controls is calculated:

$$EOR = \frac{D_1/D_0}{C_1/C_0} = \frac{\text{cases: exposed / unexposed}}{\text{controls: exposed / unexposed}}$$

Traditional case-noncase design:

EOR estimates the *incidence odds ratio* (OR).

Estimation of "relative risk" (cont'd)

In nested design (B) the

exposure odds C_1/C_0 among controls

is a statistically valid estimate of the

exposure odds Y_1/Y_0 of person-times

in the whole cohort.

Hence, exposure odds ratio EOR between cases and controls is a valid and efficient estimate of the unknown hazard ratio HR!

Role of controls:

- NOT to represent "non-cases"; those who remain healthy, but
- they provide data on exposure distribution in the whole cohort

Precision and efficiency

Case-control study: Variance of log(EOR)

$$= \frac{1}{D_1} + \frac{1}{D_0} + \frac{1}{C_1} + \frac{1}{C_0}$$

= cohort variance + sampling variance

- Depends basically on the numbers of cases, when there are 4 or more controls per case.
- Is not much bigger than variance in a cohort study with same numbers of cases.

⇒ Usually no need for more than 4 controls per case.

⇒ *Case-control design is very cost-efficient!*

Example: Smoking and cervical cancer

	Cotinine level (ng/ml)			Total
	> 242.6	3 – 242.6	> 3	
Cases	55	46	70	171
Controls	124	127	245	496

For illustration only! Pretend that the data were unmatched.

Crude estimate of HR btw high and low levels:
 $EOR = (55/124)/(70/245) = 1.6$

Don't do this when really analysing matched data!!

Variance of $\log(EOR) = 1/55 + 1/70 + 1/124 + 1/245$
 $= 0.032 + 0.012 = 0.044$; 95% CI: 1.0 to 2.4;

Variance increased only by $0.012/0.032 = 38\%$ with 500 controls compared to full cohort sampling with 0.5 million "controls"!

Warnings for overmatching

- Matching on an *intermediate* variable between exposure and outcome
 ⇒ *bias* in effect estimation
- Matching on a *correlate* of exposure which is not a risk factor of outcome
 ⇒ *loss of efficiency* in estimation.
- **Counter-matching:** Choose a control which is not similar to the case w.r. to a correlate of exposure.
 ⇒ increases efficiency.

Comparison of designs (cont'd)

- Missing data
 NCC: In 1:1 matching the case-control pair is lost, if either of the two has data missing on key risk factors.
 CC: Missingness of few data items is less serious.
- Quality and comparability of biological measurements
 NCC: Allows each case and its controls to be matched also for analytic batch, storage time, freeze-thaw cycle, → removes differential misclassification.
 CC: Measurements for the control group are performed at different times than for cases → quality problems and differential misclassification due to the above factors.
- Reuse of controls for new outcome diseases?
 NCC: Close matching makes this problematic: prone to bias and inefficiency.
 CC: This possibility is an inbuilt feature of the design.

Matching in nested design

- Special form of stratified sampling of controls.
- Creates similar distribution of a few important confounders (e.g. region, age, sex) in controls as among the cases.
- Comparable quality of measurements, when matching for storage time, freeze-thaw cycle, analytic batch.
- Controls confounding due to matching factors – but only if properly analyzed(!).
- Increases sometimes precision & efficiency in HR estimation (more balanced comparisons within the matched strata).
 – One of main reasons for matching.

Matching must be taken into account in the analysis by appropriate stratification & Mantel-Haenszel method, or modelling (conditional logistic regression).

Comparison of NCC and CC designs

- Statistical efficiency
 Roughly similar with same amount of cases and controls
- Statistical analysis and inference
 NCC: Straightforward with widely available software fitting conditional logistic regression or proportional hazards models.
 CC: More complicated; software for PH models can be used but using some tricks to obtain correct standard errors.
- Analysis on different time scales?
 NCC: No, limited to the time scale used in risk set definition.
 CC: Yes, different time scales can be used.

Conclusion

- Cost-efficient designs based on "case-controlling" are available and widely used in biobank-based epidemiologic research.
- The application of these designs requires adequate statistical expertise both in the planning and in the analysis stage.
- The nested case-control design is better suited for studying biomarkers that can be influenced by analytic batch, long-term storage, and freeze-thaw cycles.