

Nested case-control and case-cohort studies

Tuesday, 06 June, 2023

Esa Läärä & Martyn Plummer & Krista Fischer

Statistical Practice in Epidemiology with R
University of Tartu, Estonia
June, 2023

Points to be covered

- ▶ Outcome-dependent sampling designs a.k.a. **case-control** studies vs. **full cohort** design.
- ▶ **Nested case-control** study (NCC): sampling of controls from risk-sets during follow-up of study population.
- ▶ **Matching** in selection of control subjects in NCC.
- ▶ R tools for NCC: function `ccwc()` in `Epi` for sampling controls, and `clogit()` in `survival` for model fitting.
- ▶ **Case-cohort** study (CC): sampling a subcohort from the whole cohort as it is at the start of follow-up.
- ▶ R tools for CC model fitting: function `cch()` in `survival`

Example: Smoking and cervix cancer

Study population, measurements, follow-up, and sampling design

- ▶ Joint cohort of $N \approx 500\,000$ women from 3 Nordic biobanks.
- ▶ Follow-up: From variable entry times since 1970s till 2000.
- ▶ For each of 200 cases, 3 controls were sampled; matched for biobank, age (± 2 y), and time of entry (± 2 mo).
- ▶ Frozen sera of cases and controls analyzed for cotinine *etc.*

Main result: Adjusted OR = 1.5 (95% CI 1.1 to 2.3) for high (>242.6 ng/ml) vs. low (<3.0 ng/ml) cotinine levels.

Simen Kapeu *et al.* (2009) *Am J Epidemiol*

Example: USF1 gene and CVD

Study population, measurements, follow-up, and sampling design

- ▶ Two FINRISK cohorts, total $N \approx 14000$ M & F, 25-64 y.
- ▶ Baseline health exam, questionnaire & blood specimens at recruitment in the 1990s – Follow-up until the end of 2003.
- ▶ Subcohort of 786 subjects sampled.
- ▶ 528 incident cases of CVD; 72 of them in the subcohort.
- ▶ Frozen blood from cases and subcohort members genotyped.

Main result: Female carriers of a high risk haplotype had a 2-fold hazard of getting CVD [95% CI: 1.2 to 3.5]

Komulainen *et al.* (2006) *PLoS Genetics*

Full cohort design & its simple analysis

- ▶ **Full cohort design:** Data on exposure variables obtained for all subjects in a large study population.
- ▶ Summary data for crude comparison:

	Exposed	Unexposed	Total
Cases	D_1	D_0	D
Non-cases	B_1	B_0	B
Group size at start	N_1	N_0	N
Follow-up times	Y_1	Y_0	Y

- ▶ Crude estimation of **hazard ratio** $\rho = \lambda_1/\lambda_0$:
incidence rate ratio IR, with standard error of $\log(\text{IR})$:

$$\hat{\rho} = \text{IR} = \frac{D_1/Y_1}{D_0/Y_0} \quad \text{SE}[\log(\text{IR})] = \sqrt{\frac{1}{D_1} + \frac{1}{D_0}}.$$

- ▶ More refined analyses: Poisson or Cox regression.

Problems with full cohort design

Obtaining exposure and covariate data

- ▶ Slow and expensive in a big cohort.
- ▶ Easier with questionnaire and register data,
- ▶ Extremely costly and laborious for e.g.
 - measurements from biological specimens, like genotyping, antibody assays, *etc.*
 - dietary diaries & other manual records

Can we obtain equally valid estimates of hazard ratios etc. with nearly as good precision by some other strategies?

Yes – we can!

Estimation of hazard ratio

The incidence rate ratio can be expressed:

$$\begin{aligned} \text{IR} &= \frac{D_1/D_0}{Y_1/Y_0} = \frac{\text{cases: exposed / unexposed}}{\text{person-times: exposed / unexposed}} \\ &= \frac{\text{exp're odds in cases}}{\text{exp're odds in p-times}} = \textbf{exposure odds ratio (EOR)} \end{aligned}$$

= Exposure distribution in cases vs. that in cohort!

Implication for more efficient design:

- ▶ *Numerator*: Collect exposure data on all cases.
- ▶ *Denominator*: Estimate the ratio of person-times Y_1/Y_0 of the exposure groups in the cohort by **sampling** “control” subjects, on whom exposure is measured.

Case-control designs

General principle: Sampling of subjects from a given study population is *outcome-dependent*.

Data on risk factors are collected separately from

- (I) **Case group:** All (or high % of) the D subjects in the study population (total N) encountering the outcome event during the follow-up.
- (II) **Control group:**
 - ▶ Random **sample** (simple or stratified) of C subjects ($C \ll N$) from the population.
 - ▶ Eligible controls must be *bf* risk (alive, under follow-up & free of outcome) at given time(s).

Study population in a case-control study?

Ideally: The study population comprises subjects who would be included as cases, if they got the outcome in the study

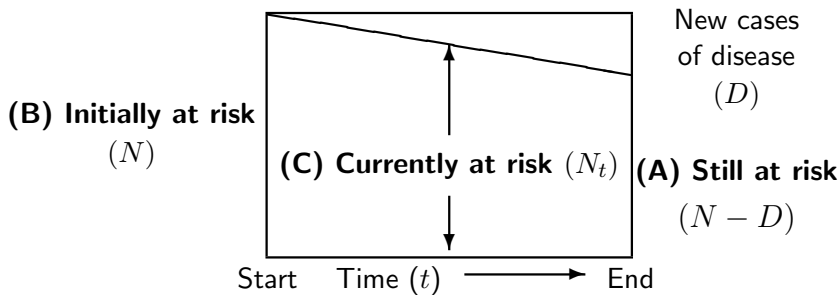
- ▶ *Cohort-based studies*: **cohort** or **closed** population of well-identified subjects under intensive follow-up for outcomes (e.g. biobank cohorts).
- ▶ *Register-based studies*: **open** or **dynamic** population in a region covered by a disease register.
- ▶ *Hospital-based studies*: dynamic **catchment** population of cases – may be hard to identify (e.g. hospitals in US).

In general, the role of control subjects is to represent the distribution of person-times by exposure variables in the underlying population from which the cases emerge.

Sampling of controls – alternative frames

Illustrated in a simple longitudinal setting:

Follow-up of a cohort over a fixed risk period & no censoring.



Rodrigues, L. & Kirkwood, B.R. (1990). Case-control designs of common diseases ... *Int J Epidemiol* **19**: 205-13.

Sampling schemes or designs for controls

(A) Exclusive or traditional, “case-noncase” sampling

- ▶ Controls chosen from those $N - D$ subjects still at risk (healthy) at the end of the risk period (follow-up).

(B) Inclusive sampling or case-cohort design (CC)

- ▶ The control group – *subcohort* – is a random sample of the cohort (N) at start.

(C) Concurrent sampling or density sampling

- ▶ Controls drawn during the follow-up
- ▶ **Risk-set or time-matched sampling:**
A set of controls is sampled from the *risk set* at each time t of diagnosis of a new case – a.k.a. **nested case-control design (NCC)**

Nested case-control – two meanings

- ▶ In some epidemiologic books, the term “nested case-control study” (NCC) covers jointly all variants of sampling: **(A)**, **(B)**, and **(C)**, from a cohort.

Rothman *et al.* (2008): *Modern Epidemiology*, 3rd Ed.

Dos Santos Silva (1999): *Cancer Epidemiology*. Ch 8-9

- ▶ In biostatistical texts NCC typically refers only to the variant of concurrent or density sampling **(C)**, in which *risk-set* or *time-matched* sampling is employed.

Borgan & Samuelsen (2003) in *Norsk Epidemiologi*

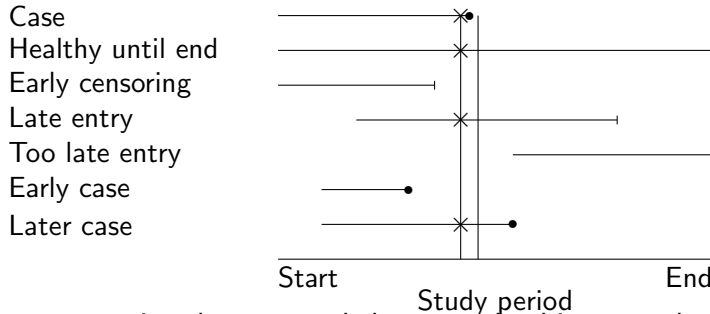
Langholz (2005) in *Encyclopedia of Biostatistics*.

- ▶ We shall follow the biostatisticians!

NCC: Risk-set sampling with staggered entry

Sampling frame to select controls for a given case:

Members (\times) of the **risk set** at t_k , i.e. the population at risk at the time of diagnosis t_k of case k .



Sampled risk set contains the case and the control subjects randomly sampled from the non-cases in the risk set at t_k .

Use of different sampling schemes

(A) Exclusive sampling, or “textbook” case-control design

- ▶ Almost exclusively(!) used in studies of epidemics.
- ▶ (Studies on birth defects with *prevalent* cases.)

(B) Inclusive sampling or case-cohort design

- ▶ Good esp. for multiple outcomes, if measurements of risk factors from stored material remain stable.

(C) Concurrent or density sampling (without or with time-matching)

- ▶ The only logical design in an open population.
- ▶ Most popular in chronic diseases (Knol *et al.* 2008).

Designs **(B)** and **(C)** allow valid estimation of hazard ratios ρ without any “rare disease” assumption.

Case-control studies: Textbooks vs. real life

- ▶ Many epi texts focus on the traditional design: **exclusive sampling** of controls, ignoring other designs.
- ▶ Claim: “*Odds ratio is the only estimable parameter.*”
- ▶ Yet, over 60% of published case-control studies apply **concurrent sampling** or **density sampling** of controls from an **open** or **dynamic** population.
- ▶ Thus, the parameter most often estimated is the **hazard ratio** (HR) or **rate ratio** ρ .
- ▶ Still, 90% of authors really estimating HR, reported as having estimated an OR (e.g. Simen Kapeu *et al.* 2009)

Knol *et al.* (2008). What do case-control studies estimate?
Am J Epidemiol **168**: 1073-81.

Exposure odds ratio – estimate of what?

- Crude summary of case-control data

	exposed	unexposed	total
cases	D_1	D_0	D
controls	C_1	C_0	C

- Depending on study base & sampling strategy, the **exposure odds ratio**

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

is a consistent estimator of

- (a) hazard ratio, (b) risk ratio, (c) risk odds ratio,
- (d) prevalence ratio, or (e) prevalence odds ratio

- **NB.** In case-cohort studies with variable follow-up times C_1/C_0 is substituted by \hat{Y}_1/\hat{Y}_0 , from estimated p-years.

Precision and efficiency

With exclusive **(A)** or concurrent **(C)** sampling of controls (unmatched), the estimated variance of $\log(\text{EOR})$ is

$$\begin{aligned}\widehat{\text{var}}[\log(\text{EOR})] &= \frac{1}{D_1} + \frac{1}{D_0} + \frac{1}{C_1} + \frac{1}{C_0} \\ &= \text{cohort variance} + \text{sampling variance}\end{aligned}$$

- ▶ Depends basically on the numbers of cases, with ≥ 4 controls per case.
- ▶ Is not much bigger than $1/D_1 + 1/D_0 = \text{variance in a full cohort study with same numbers of cases.}$
- \Rightarrow Usually < 5 controls per case is enough.
- \Rightarrow *These designs are very cost-efficient!*

Estimation in concurrent or density sampling

- ▶ Assume a simple situation: Prevalence of exposure in the study population stable over time.
- ⇒ The exposure odds C_1/C_0 among controls
= a consistent estimator of exposure odds Y_1/Y_0 of person-times.
- ▶ Therefore, the crude EOR = $(D_1/D_0)/(C_1/C_0)$
= a consistent estimator of hazard ratio $\rho = \lambda_1/\lambda_0$.
- ▶ Variance of $\log(\text{EOR})$ estimated as above.
- ▶ Yet, stability of exposure distribution may be unrealistic, especially in a closed study population or cohort.
- ▶ Solution: **Time-matched** sampling of controls from **risk sets**, *i.e.* NCC, & matched EOR to estimate HR.

Prentice & Breslow (1978), Greenland & Thomas (1982).

Matching in case-control studies

- = **Stratified sampling** of controls, e.g. from the same region, sex, and age group as a given case
- ▶ **Frequency matching** or **group matching**:
For cases in a specific stratum (e.g. same sex and 5-year age-group), a set of controls from a similar subgroup.
- ▶ **Individual matching** (1:1 or 1:m matching):
For each case, choose 1 or more (rarely > 5) closely similar controls (e.g. same sex, age within ± 1 year).
- ▶ **NCC**: Sampling from risk-sets implies time-matching at least. Additional matching for other factors possible.
- ▶ **CC**: Subcohort selection involves no matching with cases.

Virtues of matching

- ▶ Increases *efficiency*, if the matching factors are both
 - (i) strong *risk factors* of the disease, and
 - (ii) *correlated* with the main exposure.
 - Major reason for matching.
- ▶ *Confounding* due to poorly quantified factors (sibship, neighbourhood, *etc.*) may be removed by close matching – only if properly analyzed.
- ▶ Biobank studies: Matching for storage time, freeze-thaw cycle & analytic batch improves **comparability of measurements** from frozen specimens
 - Match on the time of baseline measurements within the case's risk set.

Warnings for overmatching

Matching a case with a control subject is a different issue than matching an unexposed subject to an exposed one in a cohort study – much trickier!

- ▶ Matching on an *intermediate* variable between exposure and outcome.

⇒ *Bias!*

- ▶ Matching on a *surrogate* or *correlate* of exposure, which is not a true risk factor.

⇒ *Loss of efficiency.*

- **Counter-matching:** Choose a control which is not similar to the case w.r.t a correlate of exposure.

⇒ Increases efficiency!

- Requires appropriate weighting in the analysis.

Sampling matched controls for NCC using R

- ▶ Suppose key follow-up items are recorded for all subjects in a cohort, in which a NCC study is planned.
- ▶ Function `ccwc()` in package `Epi` can be used for risk-set sampling of controls. – Arguments:

- `entry` : Time of entry to follow-up
 - `exit` : Time of exit from follow-up
 - `fail` : Status on exit (1 for case, 0 for censored)
 - `origin` : Origin of analysis time scale (e.g. time of birth)
 - `controls` : Number of controls to be selected for each case
 - `match` : List of matching factors
 - `data` : Cohort data frame containing input variables

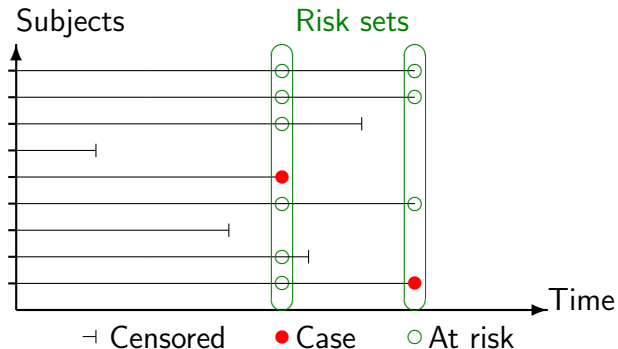
- ▶ Creates a data frame for a NCC study, containing the desired number of matched controls for each case.

Analysis of matched studies

- ▶ Close matching induces a new parameter for each matched case-control set or stratum.
 - ⇒ **unconditional logistic regression** breaks down.
- ▶ Matching on well-defined variables (like age, sex)
 - include these factors as covariates.
- ▶ Matching on “soft” variables (like sibship) can be dealt with **conditional logistic regression**.
- ▶ Same method in matched designs **(A)**, exclusive, and **(C)**, concurrent, but interpretation of β_j s differs:
 - (A)** $\beta_j = \log$ of risk odds ratio (ROR),
 - (C)** $\beta_j = \log$ of hazard ratio (HR).

Full cohort design: Follow-up & risk sets

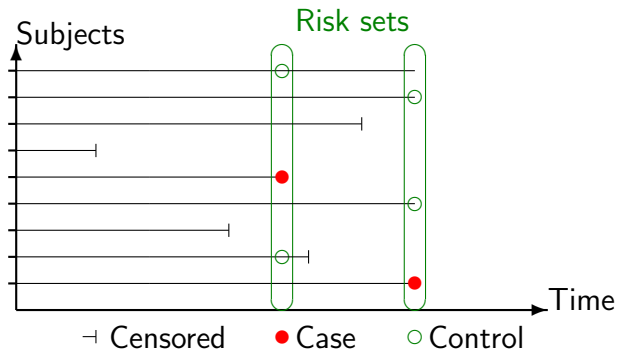
Each member of the cohort provides exposure data for all cases, as long as this member is at risk, *i.e.* (i) alive, (ii) not censored & (iii) free from outcome.



Times of new cases define the **risk-sets**.

Nested case-control (NCC) design

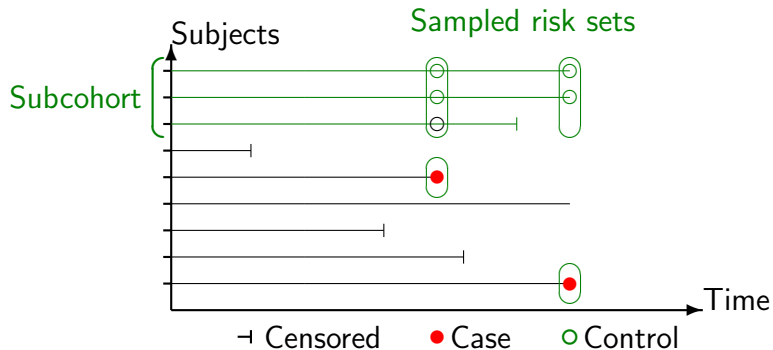
Whenever a new case occurs, a set of controls (here 2/case) are sampled from its risk set.



NB. A control once selected for some case can be selected as a control for another case, and can later on become a case, too.

Case-cohort (CC) design

- Subcohort:** Sample of the whole cohort randomly selected at the outset.
- Serves as a reference group for all cases.



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

Modelling in NCC and other matched studies

Cox proportional hazards model:

$$\lambda_i(t, x_i; \beta) = \lambda_0(t) \exp(x_{i1}\beta_1 + \cdots + x_{ip}\beta_p),$$

Estimation: partial likelihood $L^P = \prod_k L_k^P$:

$$L_k^P = \exp(\eta_{i_k}) / \sum_{i \in \tilde{R}(t_k)} \exp(\eta_i),$$

where $\tilde{R}(t_k) =$ **sampled risk set** at observed event time t_k , containing the case + sampled controls ($t_1 < \cdots < t_D$)

\Rightarrow Fit stratified Cox model, with $\tilde{R}(t_k)$'s as the strata.

\Leftrightarrow **Conditional logistic regression**

– function `clogit()` in `survival`, wrapper of `coxph()`.

Modelling case-cohort data

Cox's PH model $\lambda_i(t) = \lambda_0(t) \exp(\eta_i)$ again, but ...

- ▶ Analysis of survival data relies on the theoretical principle that *you can't know the future*.
- ▶ Case-cohort sampling breaks this principle: cases are sampled based on what *is known* to be happening to them during follow-up.
- ▶ The union of cases and subcohort is a mixture
 1. random sample of the population, and
 2. “high risk” subjects who are *certain* to become cases.

⇒ Ordinary Cox partial likelihood is wrong.

- ▶ Overrepresentation of cases must be corrected for, by (I) **weighting**, or (II) **late entry method**.

Correction method I – weighting

The method of **weighted partial likelihood** borrows some basic ideas from survey sampling theory.

- ▶ Sampled risk sets

$\tilde{R}(t_k) = \{\text{cases}\} \cup \{\text{subcohort members}\}$ at risk at t_k .

- ▶ Weights:

- $w = 1$ for all cases (within and outside the subcohort),
- $w = N_{\text{non-cases}}/n_{\text{non-cases}} = \text{inverse of sampling-fraction } f \text{ for selecting a non-case to the subcohort.}$

- ▶ Function `coxph()` with option `weights = w` would provide consistent estimation of β parameters.

- ▶ However, the SEs must be corrected!

- ▶ R solution: Function `cch()` – a wrapper of `coxph()` – in package `survival`, with `method = "LinYing"`.

Comparison of NCC and CC designs

- ▶ Statistical efficiency

Broadly similar in NCC and CC with similar numbers of cases and controls.

- ▶ Statistical modelling and valid inference

Straightforward for both designs with appropriate software, now widely available for CC, too

- ▶ Analysis of outcome rates on several time scales?

NCC: Only the time scale used in risk set definition can be the time variable t in the baseline hazard of PH model.

CC: Different choices for the basic time in PH model possible, because subcohort members are not time-matched to cases.

Comparison of designs (cont'd)

- ▶ Missing data

NCC: With close 1:1 matching, a case-control pair is lost, if either of the two has data missing on key exposure(s).

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, → better comparability.

CC: Measurements for subcohort performed at different times than for cases → differential quality & misclassification.

- ▶ Possibility for studying many diseases with same controls

NCC: Complicated, but possible if matching is not too refined.

CC: Easy, as no subcohort member is “tied” with any case.

Conclusion

- ▶ “Case-controlling” is very cost-effective.
- ▶ Case-cohort design is useful especially when several outcomes are of interest, given that the measurements on stored materials remain stable during the study.
- ▶ Nested case-control design is better suited *e.g.* for studies involving biomarkers that can be influenced by analytic batch, long-term storage, and freeze-thaw cycles.
- ▶ Matching helps in improving efficiency and in reducing bias – but only if properly done.
- ▶ Handy R tools are available for all designs.