6.1 Extension to nested case-control

Case-cohort designs

We do not need the whole cohort to estimate HR

We have seen that instead of the full risk set at each event time, we can represent each risk set by a sub-sample of the “at-risk” individuals at that time point (i.e., concurrent sampling)

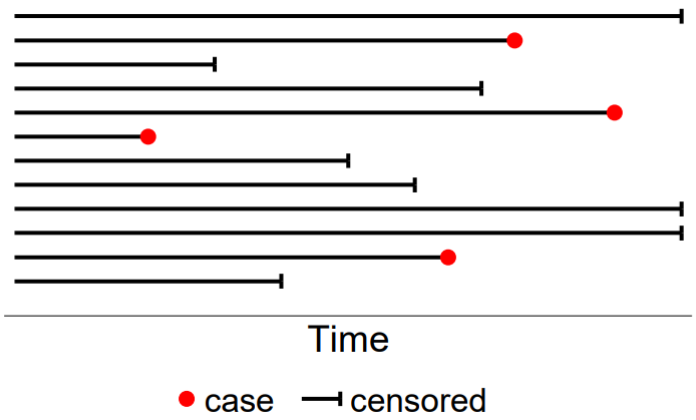
This is the nested case-control design

But we could also represent the experience of the whole cohort from a representative subsample from baseline (i.e., inclusive sampling)

This is the idea of the case-cohort design

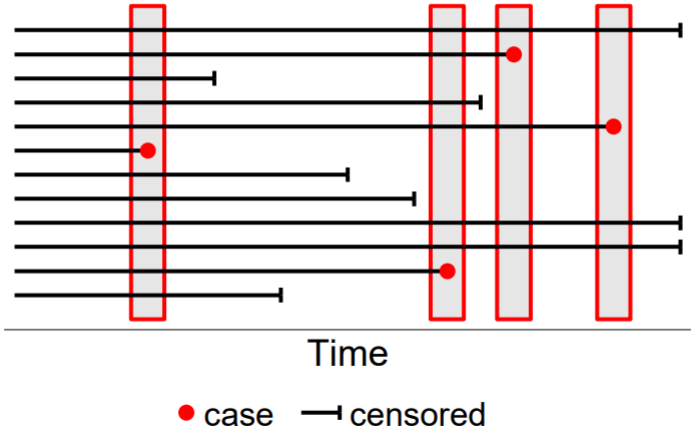
Full cohort

Cohort



Risk set (R_i)

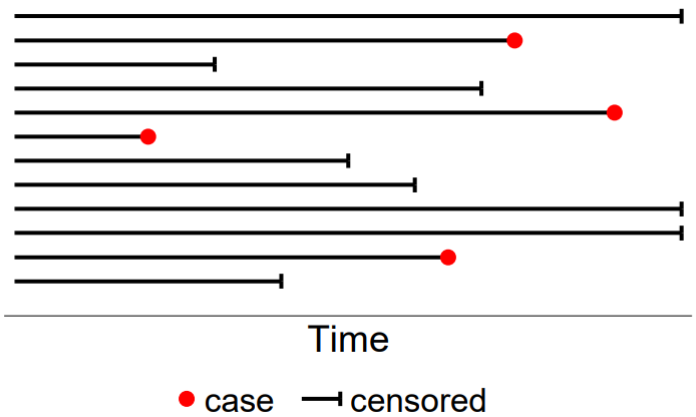
R_i={●, -, …, -}



$$L(\beta,\gamma) = \prod_{t_i} \frac{\exp^{\beta X_i + \gamma Z_i}}{\sum_{k \epsilon R_i} \exp^{\beta X_k + \gamma Z_k}}$$

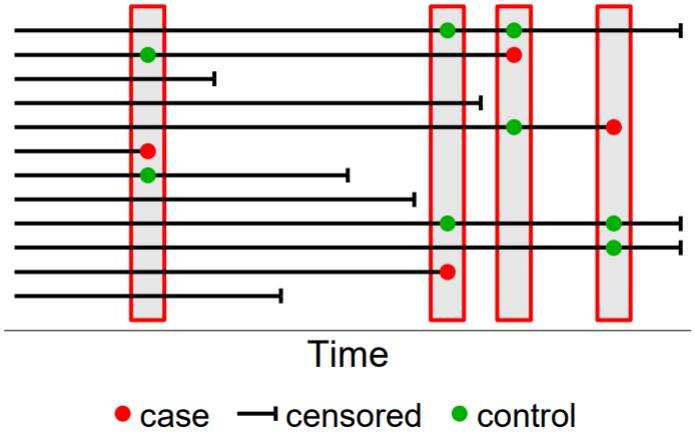
Nested case-control (NCC)

Cohort



NCC: Risk set (R_i^*)

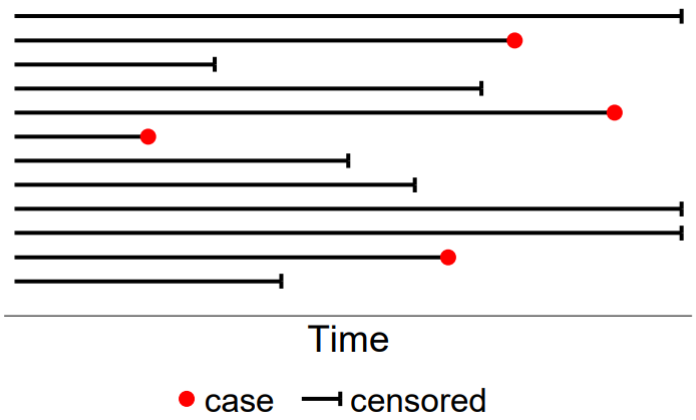
$R_i^* = \{\bullet, \bullet, \bullet\}$



$$L(\beta,\gamma) = \prod_{t_i} \frac{\exp^{\beta X_i + \gamma Z_i}}{\sum_{k \epsilon R_i^*} \exp^{\beta X_k + \gamma Z_k}}$$

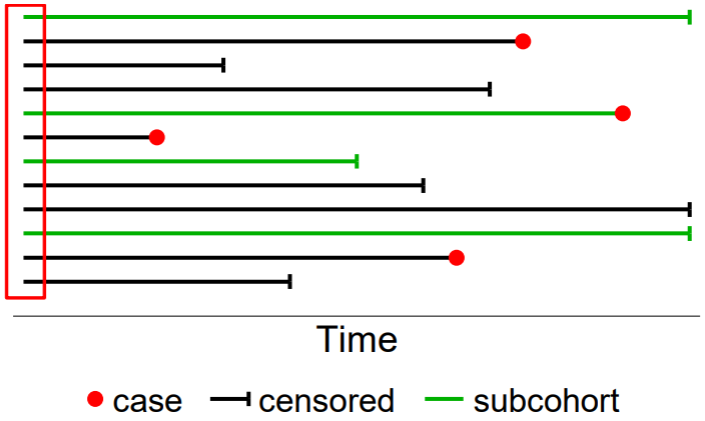
Case-cohort (CCH)

Cohort



CCH: Risk set ($S_i^{\#}$)

$S_i^{\#}$={●, -/-, …, -/-}



$$L(\beta,\gamma) = \prod_{t_i} \frac{\exp^{[\beta X_i + \gamma Z_i]}}{\sum_{k \epsilon S_i^{\#}} \exp^{[\beta X_k + \gamma Z_k]} w_k}$$

Case-cohort (CCH) or nested case-control (NCC)

Outcomes

- **NCC:** Controls used for one specific outcome because of the time-matching (i.e., concurrent sampling)
- **CCH:** The sub-cohort could be used to study multiple outcomes (i.e., inclusive sampling)

Exposures

- **NCC:** Not suitable when exposure is rare or time-varying
- **CCH:** Suitable for rare or time-varying exposures

Missing data

- **NCC:** Missing exposure/confounder variable for a control in a 1:1 study results in loss of risk set
- **CCH:** Only observation with missing data is lost from analysis

Case-cohort (CCH) or nested case-control (NCC)

Extended follow-up

- **NCC:** New cases may need new controls to be identified, enrolled and measured
- **CCH:** No new enrolments necessary with additional cases as same sub-cohort can be used for extended follow-up time

Data collection

- **NCC:** Information on cases and controls obtained at the same time, requiring constant effort/time throughout follow-up
- **CCH:** The sub-cohort is identified at the start of follow-up, so data collection can start immediately and be conducted in a short time (e.g., acute outcomes)

Case-cohort (CCH) or nested case-control (NCC)

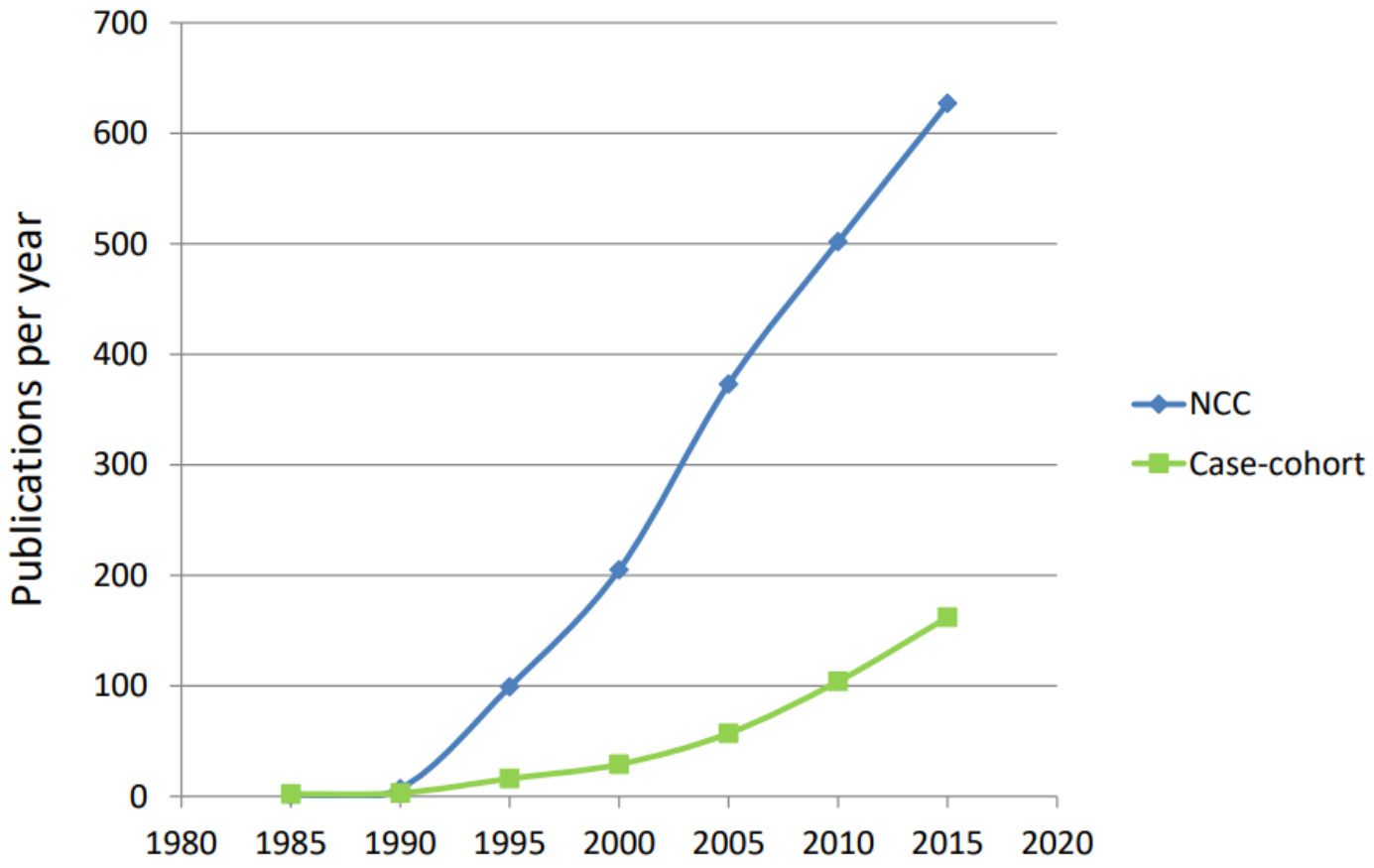
Risk measures

- **NCC:** Only a relative risk measure (i.e., hazard ratio [HR]) can be estimated from the matched data
- **CCH:** Besides HR, it is possible to estimate the prevalence, relative risk (RR) and cumulative incidence

Model

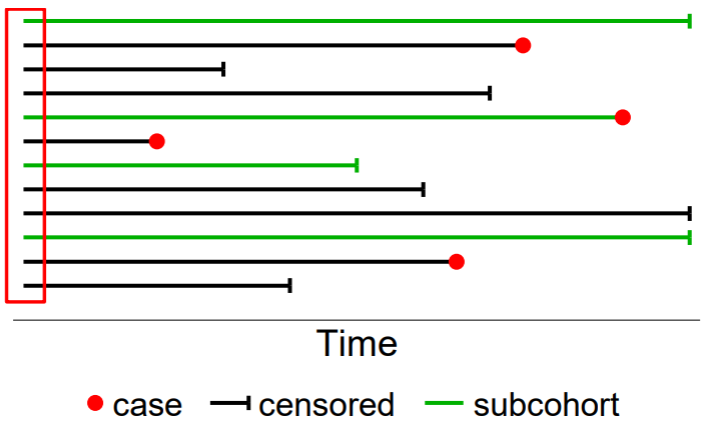
- **NCC:** Matched sets analyzed by conditional logistic regression (or logistic regression if stratum size is large)
- **CCH:** Flexible with respect to model used and method of analysis

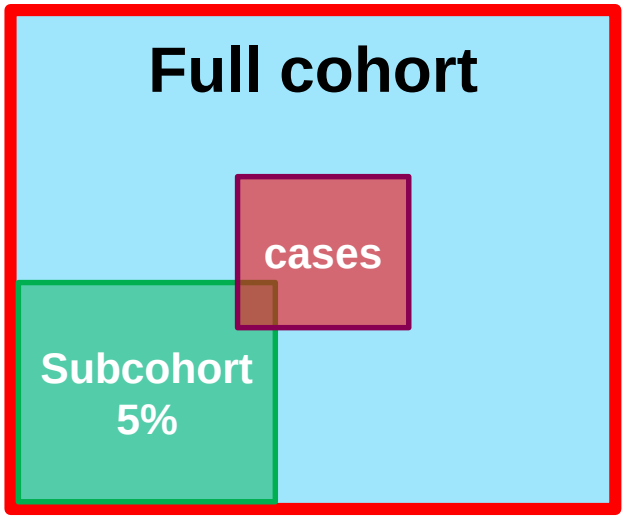
References to nested case-control and case-cohort in Web of Science



Case-cohort design

- From the cohort, select a sub-cohort of individuals at start of follow-up
- All cases that occur outside the sub-cohort during follow-up are sampled



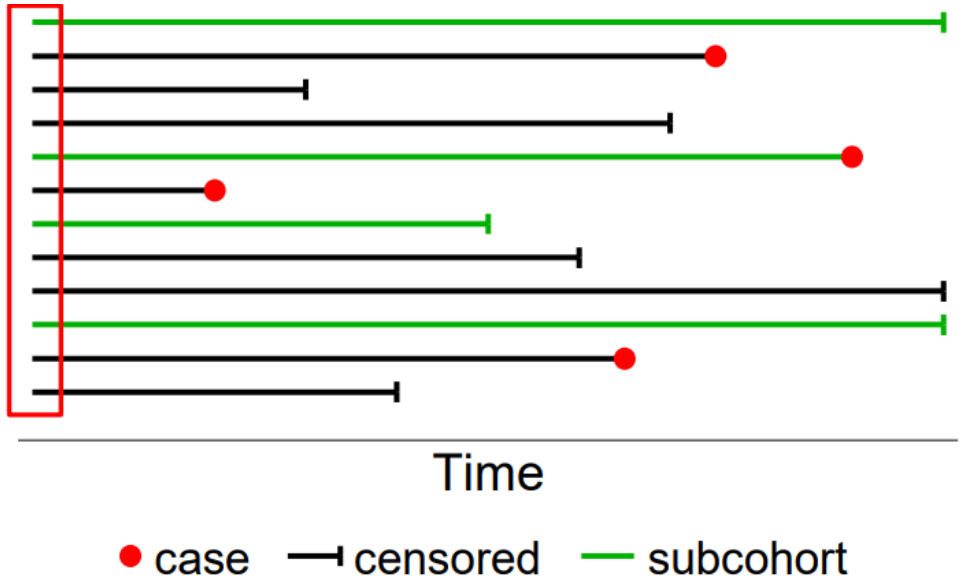


- Final sample consists of

  Sub-cohort at baseline + cases outside sub-cohort

Case-cohort design: the concept

Random subset
n% (often ~5%)
of the individuals
at baseline
AND
include ***all*** cases



Some sub-cohort members may later become cases
Cases not sampled in the sub-cohort are all included

- Information about population at risk is available in the sub-cohort+cases
- HR can be estimated and also hazard

Prentice partial likelihood:

Cox model: $h(t|X,Z) = h_0(t)\exp^{\beta X+\gamma Z}$

Cohort:

$$\prod_{t_i} \frac{\exp^{\beta X_i+\gamma Z_i}}{\sum_{k\epsilon R_i} \exp^{\beta X_k+\gamma Z_k}}$$

Full risk set

Case-cohort:
(Prentice Likelihood)

$$\prod_{t_i} \frac{\exp^{\beta X_i+\gamma Z_i}}{\sum_{k\epsilon S_i^{\#}} \exp^{\beta X_k+\gamma Z_k}}$$

Subcohort and the case at risk at time t_i

- Cases over-represented requiring ”reweight” to correct for biased sampling
- Variance for the same control population is upweighted and used repeatedly over time, resulting in biased variance requiring adjustment

Modification to Prentice likelihood

Different schemes proposed involve:

including “future” cases at times prior to their event

weighting available data to best represent the cohort

Good overview in Kulathinal (2007)

Overview of main idea:

- Each observation is given a weight, pending on case or non-case status
- Based on theory of inverse probability weighting (IPW)
- Weighted likelihood is a pseudo-likelihood which is used to estimate parameters and obtain confidence intervals
- Correct standard error (SE) by using robust SE (e.g., sandwich estimator) because pseudo-likelihood is upweighting the same individuals

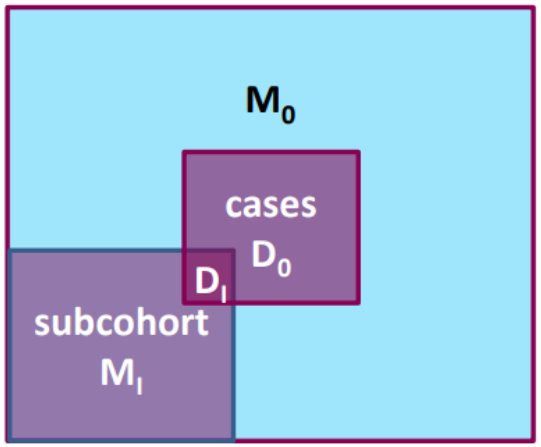
By weighting the case-cohort data, we represent the full cohort!

To compute weights, we need to keep track of numbers of cases/non-cases in/outside the sub-cohort

Kulathinal S, Karvanen J, Saarela O, Kuulasmaa K. Case-cohort design in practice - experiences from the MORGAM Project. Epidemiol Perspect Innov. 2007;4:15

Keep track of numbers

	Outside subcohort	Inside subcohort	Total
Non-case	M_0	M_I	M
Case	D_0	D_I	D
Total	$\mathbf{N_0}$	$\mathbf{N_I}$	**N**



- Sampling fraction: $p = \frac{N_I}{N}$
- Sampling fraction non-cases: $p_M = \frac{M_I}{M} \approx p$
- Sampling fraction cases: $p_D = \frac{D_0 + D_I}{D} = 1$

When full cohort is enumerated, M_0, M_I, D_0 and D_I are known.
Exposure will be known for M_I, D_0 & D_I.

Case-cohort analysis: weighted likelihood

Cox model: $h(t|X,Z) = h_0(t)\exp^{\beta X+\gamma Z}$

Cohort:

$$\prod_{t_i} \frac{\exp^{\beta X_i+\gamma Z_i}}{\sum_{k\in R_i} \exp^{\beta X_k+\gamma Z_k}}$$

risk set

Case-cohort:

$$\prod_{t_i} \frac{\exp^{[\beta X_i+\gamma Z_i]}}{\sum_{k\in S_i^{\#}} \exp^{[\beta X_k+\gamma Z_k]} w_k}$$

Joint set of the subcohort and all cases at risk at time t_i

weight for subject k

Weighted likelihood approach

Previous slide was **<u>Borgan II weights</u>** [Borgan et al, 2000]

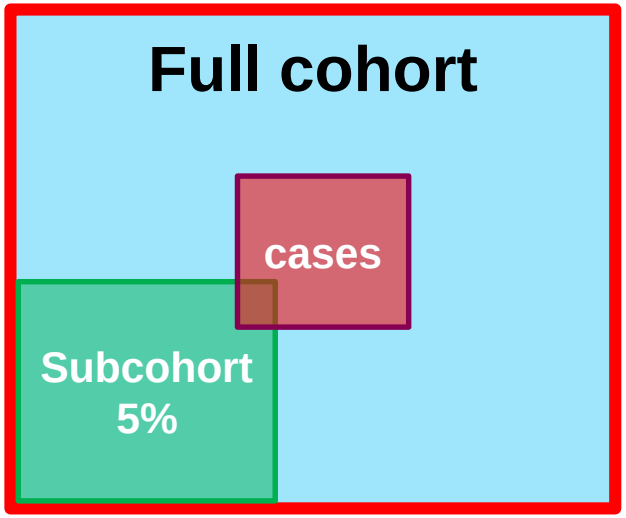
The case-cohort sample contains
all cases in the cohort :

→ Each case has weight = 1 in the analysis

The case-cohort sample contains a
subset of the cohort's non-cases:

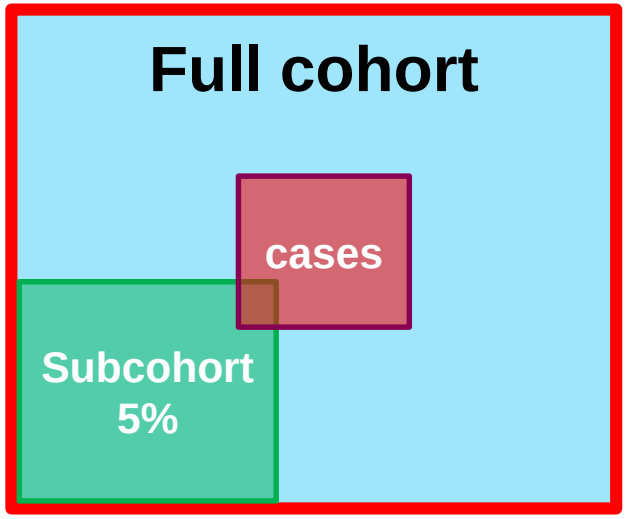
→ Each non-case has weight $w=1/p_M$

(p_M =sampling fraction of non-cases)



Example (Swedish population data)

- Swedish women born 1948-1952 in MGR (full cohort)
  - → Breast cancers occurring in ages 25-50 years.
  - → N=323,850
  - → Defined cohort, follow-up times for women



- **Sampling of case-cohort** design:
  - → A subcohort of 5% were randomly drawn
  - → All breast cancer cases occurring outside the subcohort were included.

- Modelling **educational level** (high vs low) as the only covariate.
  - → Full cohort and case-cohort
  - → Cox model using Borgan II weights (weighted approach)

Sampling Fractions

case	subcoh 0	subcoh 1	Total
0	302,939	15,990	318,929
1	4,692	229	4,921
Total	307,631	16,219	323,850

non-cases:

$$p_M = \frac{15,990}{318,929} = 0.050137$$

total:

$$p = \frac{16,219}{323,850} = 0.050082$$

Full cohort: n= 323,850

Case-cohort: n= 20,911 (15,990 + 4,692+229)

Results: Education level and breast cancer

		Cox Model
Full cohort	HR	0.8363
	β SE	-0.1787 0.0318
Case-cohort (Borgan II)	HR	0.8270
	β SE*	-0.1900 0.0358

similar (sampling variation may cause some difference)

Full cohort n=323,850, cases n=4,921

Case-cohort n=20,911, cases n=4,921

*Robust SE

Results: Education level and breast cancer

		Cox Model
Full cohort	HR	0.8363
	β SE	-0.1787 0.0318
Case-cohort (Borgan II)	HR	0.8270
	β SE*	-0.1900 0.0358

additional error very small vs. gain in dataset reduction.

Full cohort n=323,850, cases n=4,921

Case-cohort n=20,911, cases n=4,921

*Robust SE

Results: Education level and breast cancer

		Cox Model	Flexible Parametric Model
Full cohort	HR	0.8363	0.8363
	β SE	-0.1787 0.0318	-0.1787 0.0318
Case-cohort (Borgan II)	HR	0.8270	0.8270
	β SE*	-0.1900 0.0358	-0.1900 0.0358

Full cohort n=323,850, cases n=4,921

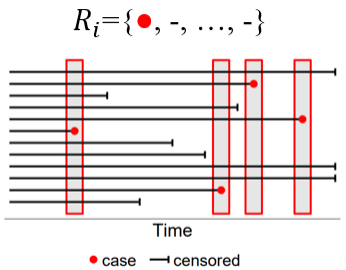
Case-cohort n=20,911, cases n=4,921

*Robust SE

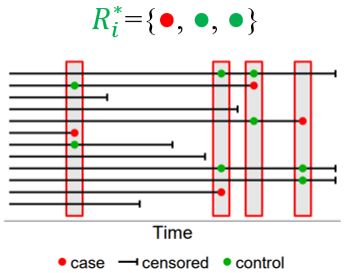
Cox and FPM are similar

The 3 partial likelihoods

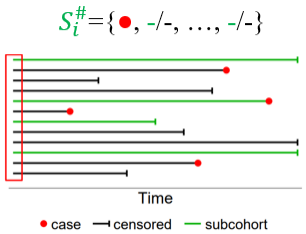
Cohort: $L(\beta,\gamma) = \prod_{t_i} \frac{\exp^{\beta X_i + \gamma Z_i}}{\sum_{k \epsilon R_i} \exp^{\beta X_k + \gamma Z_k}}$



NCC: $L(\beta,\gamma) = \prod_{t_i} \frac{\exp^{\beta X_i + \gamma Z_i}}{\sum_{k \epsilon R_i^*} \exp^{\beta X_k + \gamma Z_k}}$



CCH: $L(\beta,\gamma) = \prod_{t_i} \frac{\exp^{[\beta X_i + \gamma Z_i]}}{\sum_{k \epsilon S_i^\#} \exp^{[\beta X_k + \gamma Z_k]} w_k}$



Borgan Norsk Epi 2003, Vandenbroucke&Pearce Int J Epi 2012 (concepts)

Summary: case-cohort design

Methodology long known

but not widely used.

→ Thought to be complicated

→ Software was not available

Kim et al. (2015) performed simulation

- rarely any notable difference between the nested case-control design analyzed with conditional logistic regression and the case-cohort design using weighted Cox regression.
- when the predictor of interest was binary, the standard case-cohort methods were often more powerful than nested case-control design analyzed with conditional logistic regresioin.

Kim RS. A new comparison of nested case-control and case-cohort designs and methods. Eur J Epidemiol. 2015;30(3):197-207.

Summary: case-cohort design

Advantages

- Same sub-cohort can be used for several outcomes
- Sub-cohort measurements at baseline (biological specimens)
- Time-scale choice flexible

Disadvantages

- Sub-cohort members that are followed rigorously have potential for being biased as representatives of the full cohort
- Changes over time in the methods of measurement used for the cases
- Sub-cohort becomes ‘thin’ latter in follow-up (e.g., censoring) resulting in some events for which there are no controls

Situations when the case-cohort design is useful

- Expensive data collection on exposures or multiple endpoints
- Reduce analytical dataset for computational efficiency (Big Data era)