Cohort analysis, risk sets, the nested case-control and case-cohort designs

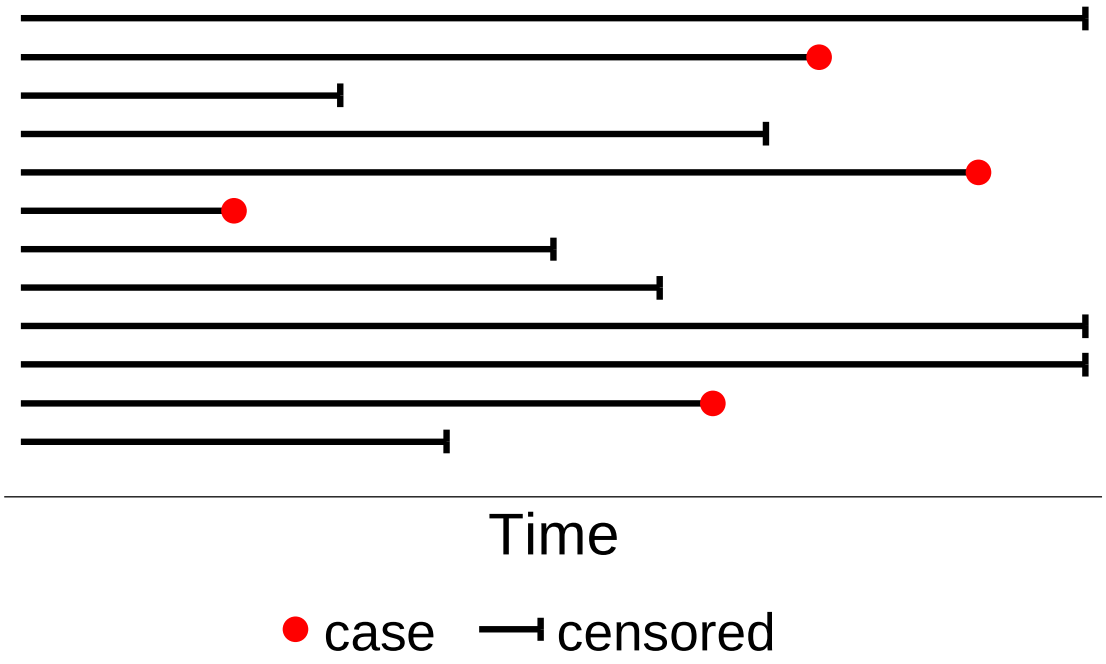
Anna Johansson
Department of Medical Epidemiology and Biostatistics (MEB)
Karolinska Institutet, Stockholm, Sweden
anna.johansson@ki.se

Biostat 3 – 2020-11-18

Aim of this lecture

- To give an introduction to risk set analysis of **cohort studies**
  - Cox regression
- To give an introduction to risk set sampling and analysis of **nested case-control** (NCC) design
  - Stratified Cox regression (conditional logistic regression)
- To give an introduction to sampling and analysis of **case-cohort** studies
  - Weighted Cox regression
- To **compare similarities and differences** between these study designs, in terms of sampling, disease measures, analysis and statistical efficiency

Cohort study



Cohort study

- The **cohort study** is characterised by
  - A group of individuals which are followed up for a specific outcome
  - Cohort members are assumed to be free of disease (outcome) at start of follow-up
  - Cohort members are followed until they have the outcome or they are censored (no longer under follow-up/observation)
  - Common reasons for censoring are
    - Death (if death is not the event of interest)
    - Emigration
    - End of study (calendar date)
    - Lost-to-follow-up

- Each cohort member contributes with **risktime** to the **cohort study base**
  - Risktime, time of follow-up, time-to-event, person-time are different names for the same thing.
  - Risktime is the time between start of follow-up and end of follow-up

Cohort study: What can we estimate (non-parametric estimation)

- From a cohort study we can calculate
  - Total **person-years at risk (Y)**, summing the risktime of all individual cohort members
  - Number of **events (D)**, that occurs in this total sum of risktime
  - The **rate λ=D/Y** is a measure of the "risk" of the outcome in the cohort; it takes into account that cohort members are followed for different lengths of time

- We are often interested in assessing the effect of an exposure on the outcome
  - Cohort members can be divided into "exposed" and "unexposed"
  - Strictly, **exposed risktime** and **unexposed risktime**

	Exposed	**Unexposed**
Cases (events)	D_1	D_0
Personyears (risktime)	Y_1	Y_0

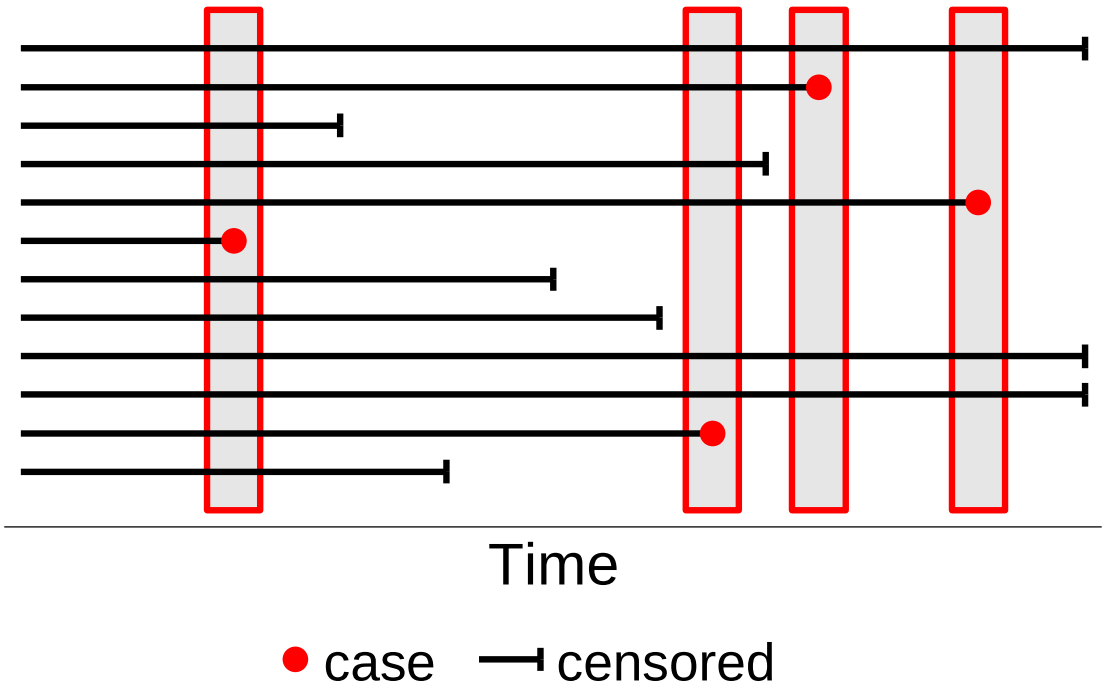
  - We can estimate the **rate of the exposed** D_1/Y_1 and **unexposed** D_0/Y_0
  - The **rate ratio,** $\lambda_1/\lambda_0=(D_1/Y_1) / (D_0/Y_0)$, is a measure of the association between the exposure and the outcome

Cohort study: Cox regression (model estimation)

- The rate can be modelled using Cox regression: $\boldsymbol{\lambda(t|X)= \lambda_0(t) \times \exp(\beta X)}$
- The parameters **β** are estimated using maximum likelihood estimation.
- Maximum likelihood method – in brief!:
  - Assume a statistical model for the data (and sometimes a distribution for the outcome).
  - The **likelihood** is the probability of the data under the model: Each observation contribute with a probability and all those probabilities are multiplied together: $L(\beta)=P_1 \times P_2 \times P_3 \times P_4 \times .....$
  - The **likelihood** is a function of the parameters of the given model and the underlying data.
  - The likelihood function is unique to each dataset.
  - We maximize the likelihood function to find the parameter values β that best describes our data, i.e. the most likely parameters.

- The likelihood for the Cox regression model is called a ”partial likelihood”, and can be used as a likelihood and maximized to obtain parameter estimates (Cox, 1972).
- It is partial because it does not include the baseline hazard part of the model, only the relative rates, **exp(βX)**.
- The Cox partial likelihood is created from risk sets.

Cohort study: Cox regression risk sets



Cohort study: Cox regression likelihood

- In each risk set, we have one event, and all other persons still at risk.
- For each event, we calculate the probability that we got that specific event in that risk set. It turns out we can use the hazards for this calculation.
- E.g. In a risk set with five persons at risk, the probability that person 2 is the event is:

$$\frac{\lambda_2}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5}$$

- Since $\boldsymbol{\lambda(t|X) = \lambda_0(t) \times \exp(\beta X)}$ and the $\boldsymbol{\lambda_0}$ cancel out, we can write this as:

$$\frac{\lambda_0 \exp(\beta x_2)}{\lambda_0 \exp(\beta x_1) + \lambda_0 \exp(\beta x_2) + \lambda_0 \exp(\beta x_3) + \lambda_0 \exp(\beta x_4) + \lambda_0 \exp(\beta x_5)} = \frac{\exp(\beta x_2)}{\sum_{i \in R} \exp(\beta x_i)}$$

- This was for one risk set. The likelihood is the product of all probabilities for all risksets (i.e. for all events).

Cohort study: Cox regression likelihood

- If we have *k* distinct event times (=all risk sets), then the partial likelihood **L(β)** is

$$L(\beta)=\prod_{j=1}^{k}\frac{\exp(\beta x_j)}{\sum_{i\in Rj}\exp(\beta x_i)}$$

- Note that these calculations do not depend on the underlying event times, only the ordering of event times is important.

Cohort study: Cox regression (R)

- Example Colon cancer, localised (stage=1), cause-specific survival (status=1)

```
> colon$event <- (colon$status==1)  ## creates indicator 0/1
> localised <- subset(colon, stage == 1)
> summary(coxph(Surv(surv_mm,event) ~ sex + factor(agegrp) + year8594,
                data = localised, ties="breslow" ))

n= 6274, number of events= 1734

                    coef exp(coef) se(coef)      z Pr(>|z|)
sex              -0.08871   0.91511  0.04937 -1.797   0.0723 .
factor(agegrp)1 -0.05217   0.94917  0.13845 -0.377   0.7063
factor(agegrp)2  0.29155   1.33850  0.12573  2.319   0.0204 *
factor(agegrp)3  0.81025   2.24848  0.12607  6.427 1.30e-10 ***
year8594         -0.28121   0.75487  0.04937 -5.696 1.23e-08 ***
```

```
se(exp(coef))
=se(coef)*exp(coef)=
=0.04937*exp(-0.08871)=
=0.045176
```

- When we fit a Cox model, the partial likelihood for the underlying model is maximized to produce the ”most likely” parameters (hazard ratios) for our data.

Cohort study: Cox regression (Stata)

- Example Colon cancer, localised (stage=1), cause-specific survival (status=1)

```
. use colon.dta, clear
. stset surv_mm if stage==1, failure(status==1) scale(12) id(id)
. stcox sex i.agegrp year8594
```

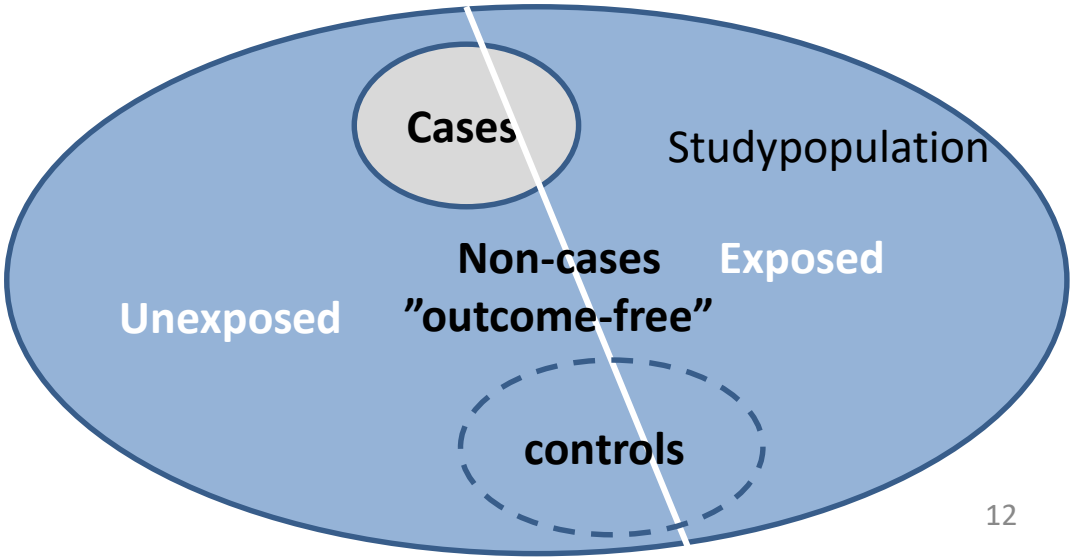
_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex	.9151101	.0451776	-1.80	0.072	.8307126	1.008082
agegrp						
45-59	.9491689	.1314101	-0.38	0.706	.723597	1.24506
60-74	1.338501	.1682956	2.32	0.020	1.046148	1.712553
75+	2.24848	.2834768	6.43	0.000	1.756199	2.878751
year8594	.7548672	.0372669	-5.70	0.000	.6852479	.8315596

- When we fit a Cox model, the partial likelihood for the underlying model is maximized to produce the ”most likely” parameters (hazard ratios) for our data.

Using a sample of controls rather than all outcome-free persons

- In situations when we are unable to, or do not want to, use a full cohort, we often consider a **case-control design**
  - Reduce the comparison group, i.e. the outcome-free group
- Reasons for case-control study: money and resources
  - Expensive data collection of exposures, e.g. genotyping or questionnaires
  - Reduce data sizes for computational efficiency, e.g. complex modelling
  - If study base is difficult to define (enumerate), e.g. no population register

Controls are randomly selected to represent the exposure distribution among non-cases



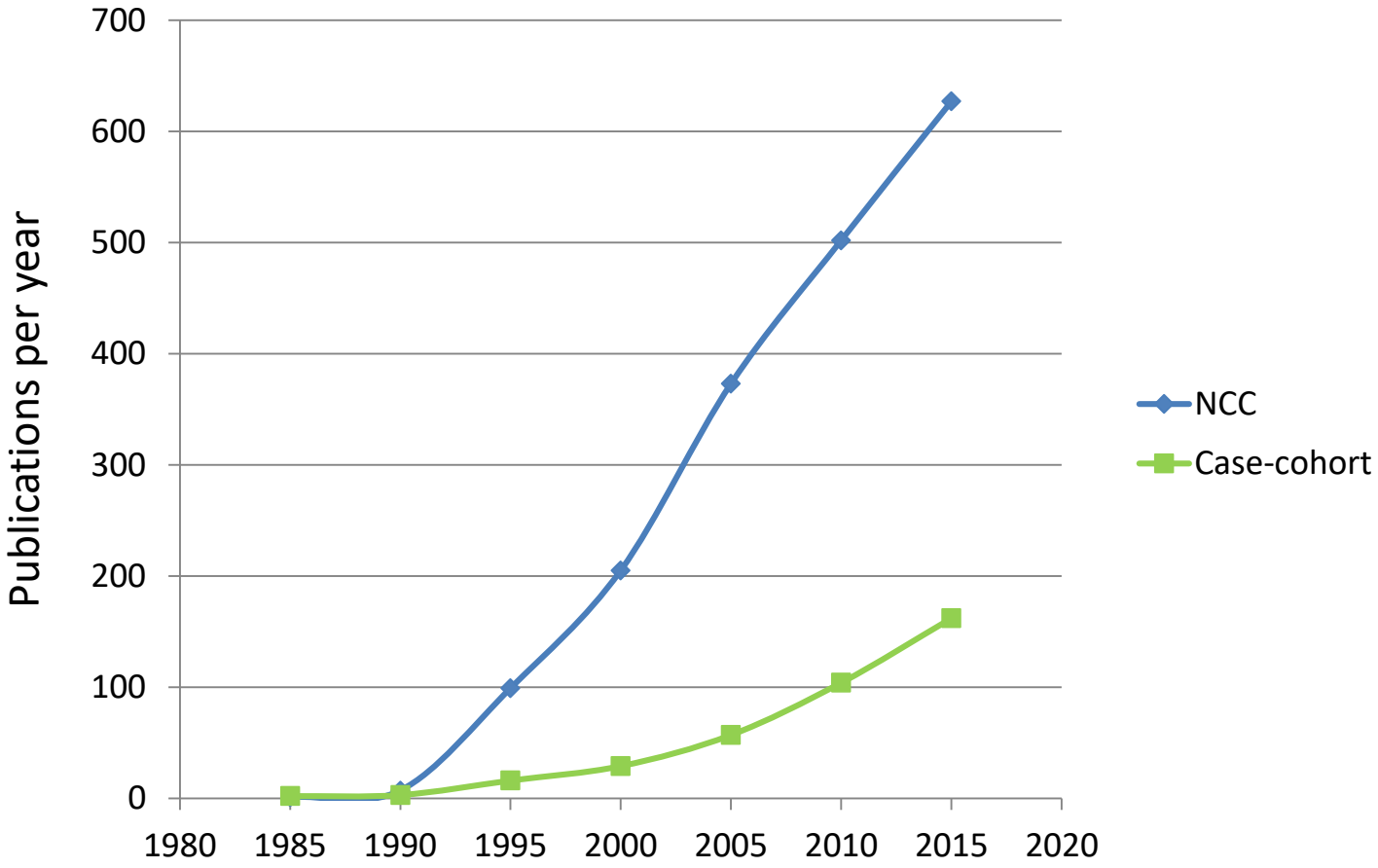
Nested case-control and Case-cohort – two design choices

- **Nested case-control design (NCC**) is the most common of the two
  - With appropriate sampling and analysis, the odds ratio from the NCC estimates the rate ratio in the full cohort. This is a key strength of the NCC!

- The **case-cohort design** is less common but gaining popularity
  - In a case-cohort study we can estimate anything that we can estimate in a cohort, including rates, rate differences and rate ratios
  - That is an advantage of the case-cohort design over the NCC design, where we typically only estimate relative measures (rate ratios) and not absolute measures (rates)
  - The rate ratio from a case-cohort design is estimating the rate ratio in the full cohort

- Why are case-cohort studies less common than NCC studies?
  - Design and analysis is thought to be complex – not true anymore!
  - Very rarely described in standard epidemiology text books!
  - Aim of this lecture is to show that case-cohort studies can be easily performed

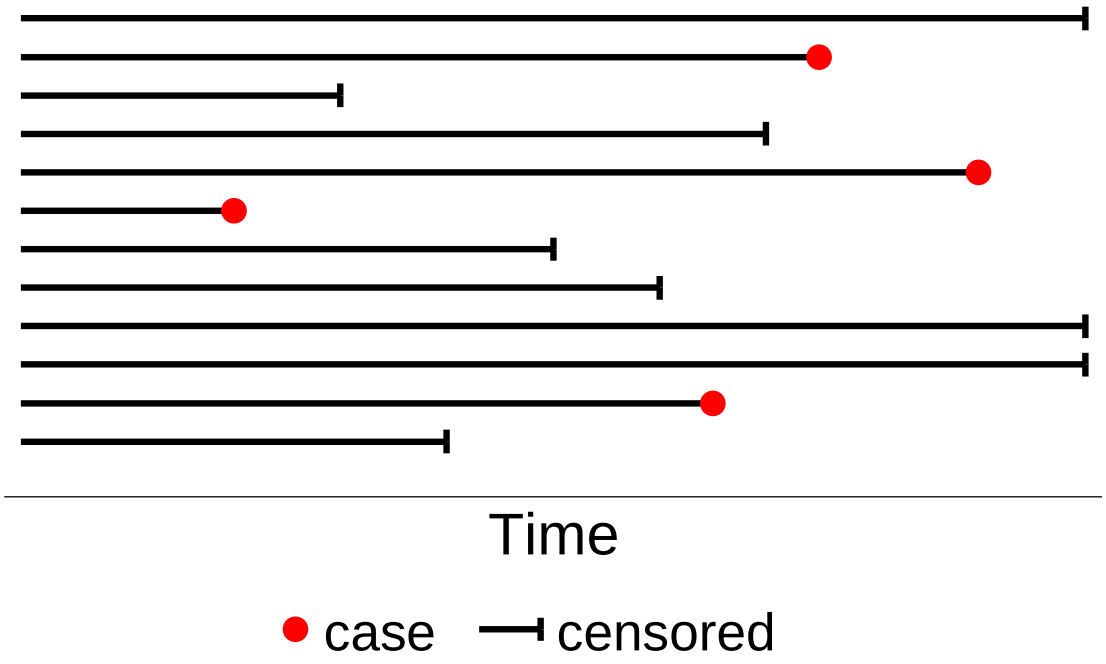
References to nested case-control and case-cohort in Web of Science



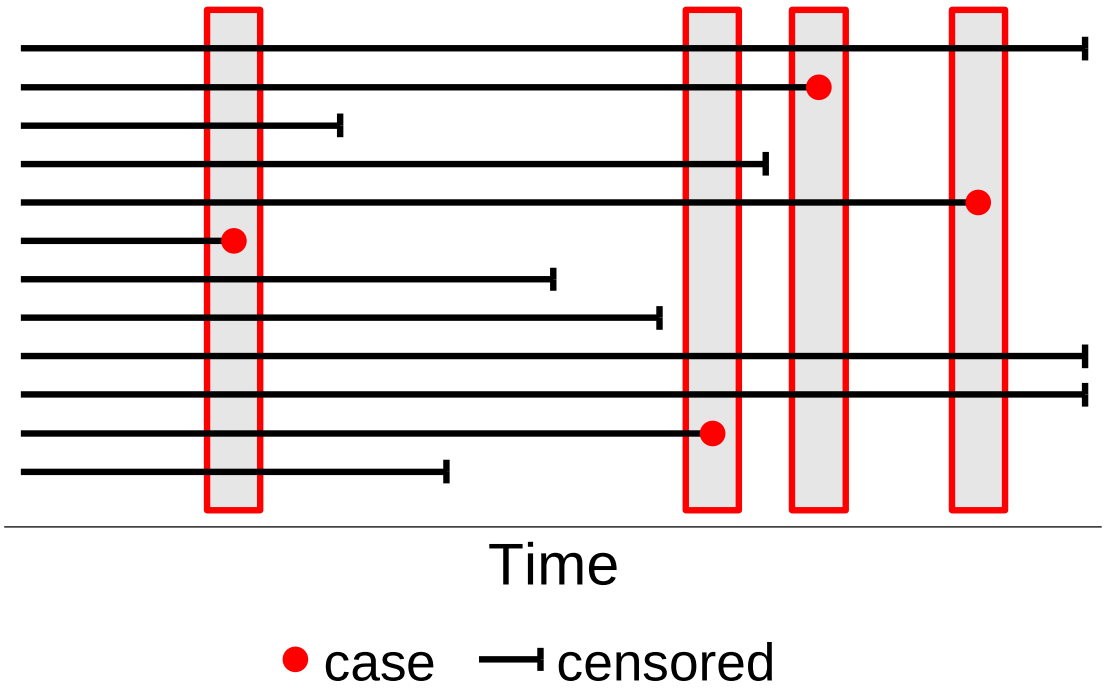
Nested Case-Control design

- We start with a cohort study….

Cohort study

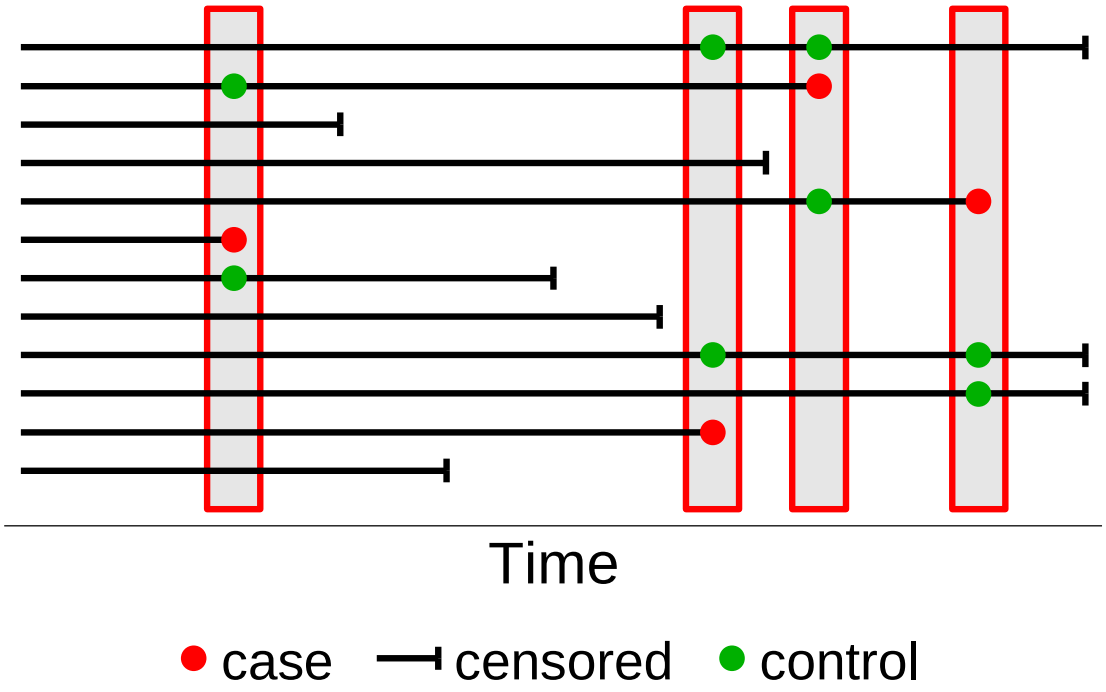


Nested Case-Control design (NCC)



Nested Case-Control design (NCC)

- If timescale is a confounder AND
- No interest in estimating effect of time



Controls are time-matched to cases.
I.e. controls can only be used for <u>one</u> outcome.

Nested Case-Control design (NCC)

- **Sampling of the NCC:**
  - Study base is some large cohort.
  - Select all those who become cases.
  - Sampling of controls ("incidence density sampling"):
    - Select controls randomly from those still at risk at time of the case ("riskset")
    - Usually 1 to 5 controls per case (>5 controls only improves statistical power minorly)
    - Controls are **time-matched** to cases. (1) Persons can be controls more than once, (2) A person selected as control may later become a case.

- Often involves additional matching on confounders.
- Analysis using conditional logistic regression, conditioning on riskset (and matching strata)
- Originally proposed by Thomas (1977), but also developed by Prentice and Breslow (1978), Oakes (1981), Goldstein and Langholz (1992)

Nested Case-Control design (NCC)

	Exposed	Unexposed
Cases	D_1	D_0
Person-years	Y_1	Y_0
Controls	C_1	C_0

- Since controls are randomly selected within risksets, the overall ratio of exposed to unexposed controls C_1/C_0 gives a consistent estimate of the ratio of exposed to unexposed pyrs Y_1/Y_0 in the cohort. This holds in general for case-control studies, and also for time matched risk set sampling.
- Hence, the (conditional) odds ratio (OR) estimates the (conditional/adjusted) rate ratio (HR)

$$HR = (D_1/Y_1) / (D_0/Y_0) = (D_1/D_0) / (Y_1/Y_0) \approx (D_1/D_0) / (C_1/C_0) = OR$$

- Controls are used instead of risktime.
- This derivation holds for constant HR (proportional hazards) (Greenland, Thomas 1982), *but I have yet to find a reference for non-proportional hazards…*
- Hence, the rare disease assumption is not required for the interpretation of the OR as an HR.

Nested Case-Control design (NCC)

- The NCC partial likelihood is very similar to the Cox partial likelihood.

$$L(\beta)=\prod_{j=1}^{k}\frac{\exp(\beta x_j)}{\sum_{i\in\tilde{R}j}\exp(\beta x_i)}$$

- $\tilde{R}j$ is the case-control sampled riskset, rather than the full riskset Rj.
- The NCC partial likelihood coincides with the conditional likelihood for matched case-control data under a logistic regression model (Prentice, Breslow 1978)
- Hence, NCC data may be analysed using conditional logistic regression, conditioning on riskset (and matching strata).
- The resulting odds ratio (OR) estimates the underlying HR in the cohort

- (Also, possible to use stratified Cox regression, stratifying on risksets and matching strata.)

Nested Case-Control design (NCC) (R)

- Generating a nested case-control study is very easy in R.
- We generate a NCC study with one control per case using **ccwc** command.

```
> set.seed(123123)      ## make sampling reproducible

## There are lots of ties, add noise to surv_mm to make them unique
> localised$surv_mm_noise <- jitter(localised$surv_mm, factor=1,
                                    amount = NULL)

> head(localised)
id sex age stage mmdx yydx surv_mm  surv_mm_noise
2    2  78     1   10 1978    82.5      82.623483
5    1  80     1    4 1980     8.5       8.525095
6    2  75     1   11 1975    23.5      23.565748
9    1  77     1    3 1977    85.5      85.463857
11   2  76     1    9 1976    32.5      32.657867
12   2  77     1    6 1977   222.5     222.410508

## Sample 1:1 control per case
> nccdata <- Epi::ccwc(exit=surv_mm_noise, fail=event, data=localised,
              include=list(sex,agegrp, year8594), controls=1, silent=TRUE)
```

Nested Case-Control design (NCC) (R)

```
## Sample 1:1 control per case
> nccdata <- Epi::ccwc(exit=surv_mm_noise, fail=event, data=localised,
              include=list(sex,agegrp, year8594), controls=1, silent=TRUE)

> tail(nccdata )
      Set  Map     Time Fail sex agegrp year8594
3463 1732 6259 8.696769    1   2      2        1
3464 1732  237 8.696769    0   2      2        0
3465 1733 6266 9.622225    1   1      3        1
3466 1733 5164 9.622225    0   1      3        1
3467 1734 6269 1.580956    1   1      3        1
3468 1734  561 1.580956    0   2      2        0

> str(nccdata )
'data.frame':   3468 obs. of  7 variables:
 $ Set     : num  1 1 2 2 3 3 4 4 5 5 ...
 $ Map     : num  2 1225 3 4863 7 ...
 $ Time    : num  8.53 8.53 23.57 23.57 36.64 ...
 $ Fail    : num  1 0 1 0 1 0 1 0 1 0 ...
 $ sex     : int  1 1 2 2 2 1 2 2 2 1 ...
 $ agegrp  : int  3 0 3 2 3 2 3 2 3 3 ...
 $ year8594: int  0 0 0 1 0 0 0 0 0 0 ...
```

Nested Case-Control design (NCC) (R)

```
## Check the data

# How many events in cohort? 1734:4540
>   table(localised$event, useNA="always")

FALSE  TRUE  <NA>
 4540  1734     0

# How many cases-controls in nccdata? 1734:1734
>   table(nccdata$Fail, useNA="always")

   0    1 <NA>
1734 1734    0
```

- After sampling we have 1734 cases and 1734 matched controls (risk set sampled).

Nested Case-Control design (NCC) (R)

- The resulting NCC study is analysed using conditional logistic regression.

```
> summary(clogit(Fail~sex + factor(agegrp)+ year8594 + strata(Set),data=nccdata))
Call:
coxph(formula = Surv(rep(1, 3468L), Fail) ~ sex + factor(agegrp) +
    year8594 + strata(Set), data = nccdata, method = "exact")

  n= 3468, number of events= 1734

                    coef exp(coef) se(coef)      z Pr(>|z|)
sex              -0.06357   0.93841  0.07093 -0.896    0.370
factor(agegrp)1 -0.12520   0.88232  0.18728 -0.669    0.504
factor(agegrp)2  0.15132   1.16337  0.17029  0.889    0.374
factor(agegrp)3  0.75649   2.13079  0.17603  4.298 1.73e-05 ***
year8594         -0.33093   0.71825  0.07176 -4.611 4.00e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Clogit is automatically transformed into stratified Cox!

- The estimates from NCC (coef(sex) -0.06357) are similar to the full cohort (coef(sex) -0.08871)
- Standard errors are slightly higher in NCC (se(sex): 0.07093) compared to the cohort (se(sex): 0.04937 cohort).

Nested Case-Control design (NCC) (Stata)

- Generating a nested case-control study is very easy in Stata.
- We generate a NCC study with one control per case using **.sttocc** command.

```
. set seed 34455667  // makes sampling reproducible
. sttocc, n(1)

          failure _d:  status == 1
    analysis time _t:  surv_mm/12
                  id:  id

There were 149 tied times involving failure(s)
  - failures assumed to precede censorings,
  - tied failure times split at random

There are 1734 cases
Sampling 1 controls for each case
..................................................................................
..........................
> ............
```

Nested Case-Control design (NCC) (Stata)

- The resulting NCC study is analysed using conditional logistic regression.

```
. clogit _case sex i.agegrp year8594, group(_set) or
                                                  Number of obs     =      3,468
                                                  LR chi2(5)        =     101.94
                                                  Prob > chi2       =     0.0000
Log likelihood = -1150.9453                       Pseudo R2         =     0.0424

------------------------------------------------------------------------------
       _case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .9058728    .063661    -1.41   0.160     .7893112    1.039648
             |
      agegrp |
      45-59  |    .927094    .168337    -0.42   0.677     .6494817    1.323368
      60-74  |   1.276786   .2123023     1.47   0.142     .9216829    1.768703
        75+  |   2.268003   .3845136     4.83   0.000     1.626793     3.16195
             |
    year8594 |   .7763301    .055581    -3.54   0.000     .6746912    .8932804
------------------------------------------------------------------------------
```

- The estimates are similar to the full cohort but standard errors are slightly higher.

Nested Case-Control design (NCC) (Stata)

- We can also analyse the NCC data using stratified Cox regression.

```
. gen surv_ncc=1 if _case==1  // make up a survival time for cases
. replace surv_ncc=2 if _case==0  // make up a survival time for controls
. stset surv_ncc, failure(_case==1)
. stcox sex i.agegrp year8594, strata(_set)

No. of subjects =        3,468                 Number of obs    =       3,468
No. of failures =        1,734
Time at risk    =         5202
                                               LR chi2(5)       =      101.94
Log likelihood  =   -1150.9453                 Prob > chi2      =      0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .9058728    .063661    -1.41   0.160     .7893112    1.039648
      agegrp |
      45-59  |    .927094    .168337    -0.42   0.677     .6494817    1.323368
      60-74  |   1.276786   .2123023     1.47   0.142     .9216829    1.768703
        75+  |   2.268003   .3845136     4.83   0.000     1.626793     3.16195
             |
    year8594 |   .7763301    .055581    -3.54   0.000     .6746912    .8932804
------------------------------------------------------------------------------
                                                        Stratified by _set
```

Nested Case-Control design (NCC)

- The log-likelihoods (-1150.9453) from conditional logistic regression and the stratified Cox model are identical, as expected, since the models are mathematically equivalent.

- Some find it difficult to grasp that a person can be both a case and a control in the same study.

- By looking at the likelihood for the Cox model – or by considering ta case-control study as being derived from a cohort study – it is obvious that it must be this way.

Nested Case-Control design (NCC)

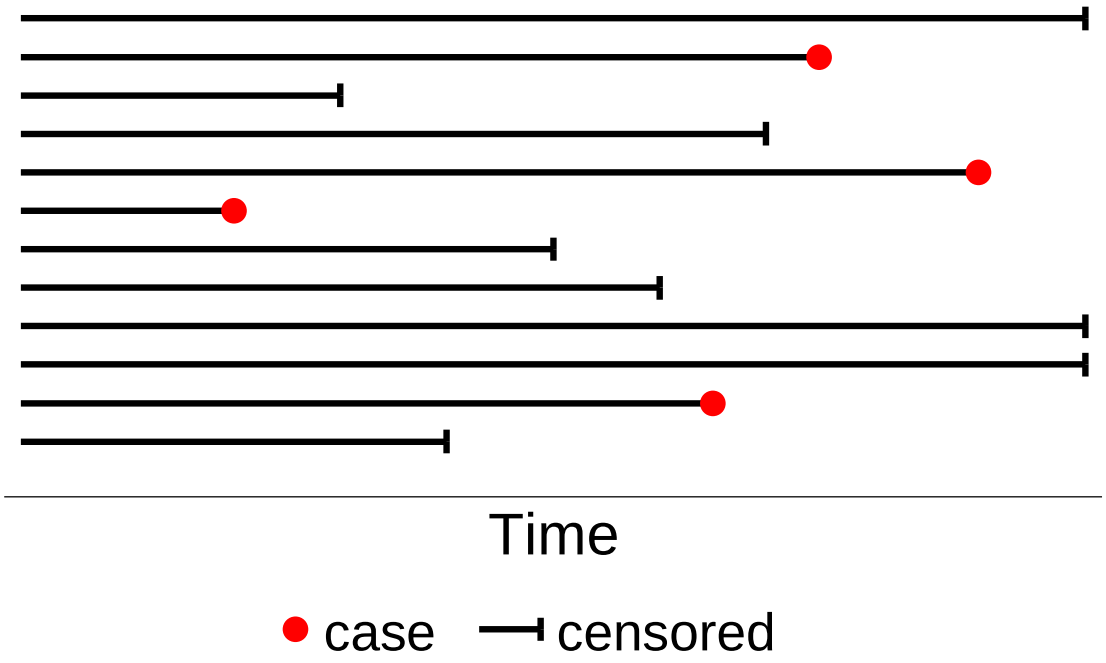
- Limitation 1:
  - The control population can only be used for **one** specific outcome (the disease that the cases have), because of the **time-matching** (incidence sampling).
  - *Not entirely true, if known sampling fractions in each riskset then controls can be re-used.*

- Limitation 2:
  - We can only estimate HRs, relative rates (rate ratios)
  - We cannot estimate rates or risks, since we do not know the underlying person-time at risk (sampling has distorted this information by selecting a fix number of controls from each riskset)
  - *Not entirely true. If we know the size of risksets and sampling fractions in each riskset, then it is possible to estimate rates (Langholz, Borgan 1997 and others). Not trivial, especially if there are time-dependent effects.*

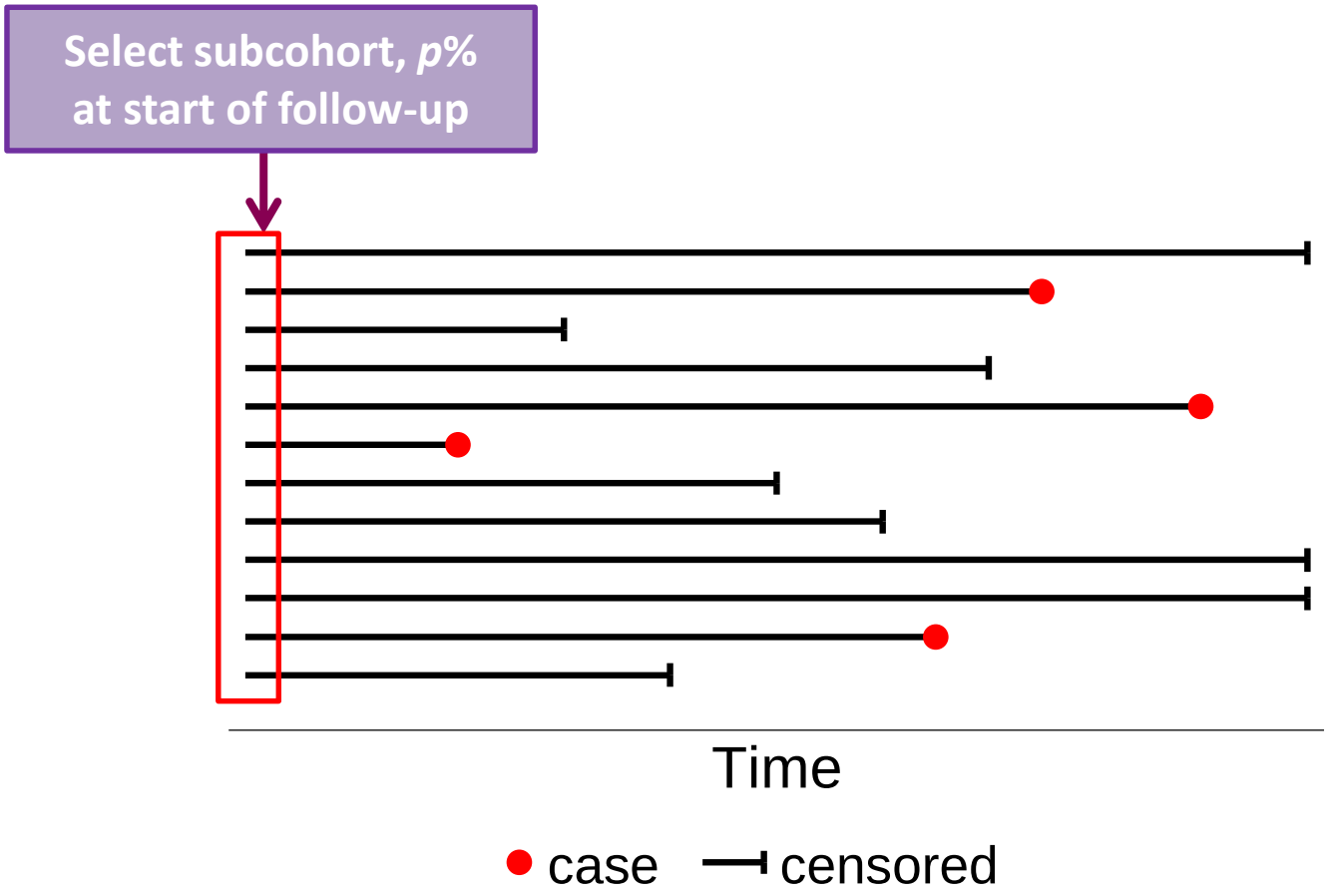
Case-cohort design

- We start with a cohort study….

Cohort study



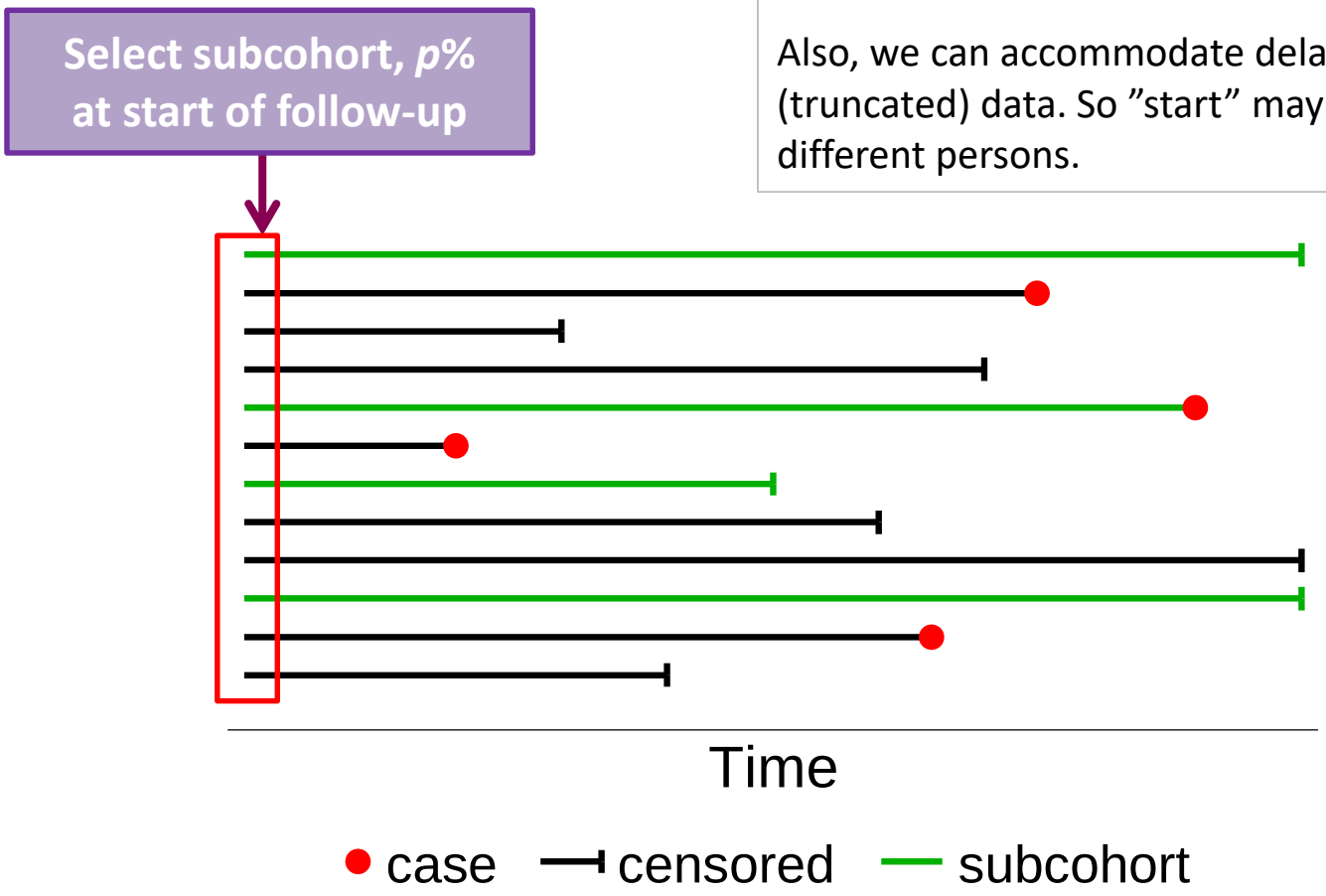
Case-cohort design



Case-cohort design

Although we say ”select at start of follow-up”, we include all persontime, i.e. selecting the ”lines” of follow-up thoughout follow-up

Also, we can accommodate delayed entry (truncated) data. So ”start” may be different for different persons.



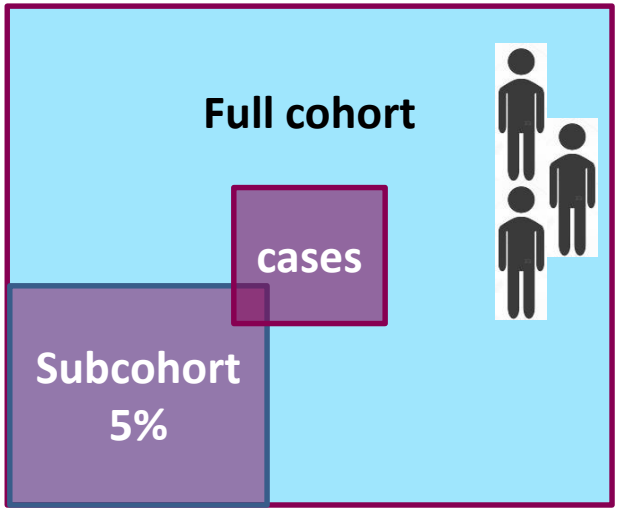
Subcohort is not time-matched to cases.
I.e. controls can be used for <u>many</u> outcomes.

Case-cohort design

- **Sampling of case-cohort:**
  - From the cohort, select a subcohort of individuals at start of follow-up.
  - The subcohort will include some cases.
  - Also include all cases that occur outside the subcohort during follow-up.
  - Final sample consists of subcohort + cases outside subcohort.

- HR can be estimated, but also hazard rates.
  - Information about population at risk is maintained via the sampling fraction

- Same subcohort can be used for several diseases (outcomes).

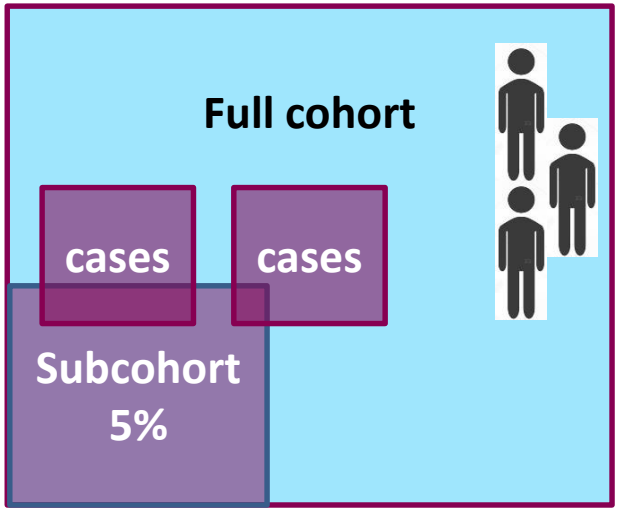


Case-cohort design

- **Sampling of case-cohort:**
  - From the cohort, select a subcohort of individuals at start of follow-up.
  - The subcohort will include some cases.
  - Also include all cases that occur outside the subcohort during follow-up.
  - Final sample consists of subcohort + cases outside subcohort.

- HR can be estimated, but also hazard rates.
  - Information about population at risk is maintained via the sampling fraction

- Same subcohort can be used for several diseases (outcomes).



Case-cohort design

- Limitation 1:
  - If many censorings, the subcohort will be "thin" in the end and not representative of the cohort. E.g. high age.
  - Reduced by stratification, with higher sampling fractions in some strata

- Limitation 2:
  - Very rarely described in any detail in standard epidemiology textbooks.
  - Good overviews can be found in Kulathinal et al 2007, Cologne et al 2012.
  - And recently: Handbook of survival analysis (2013), chapter 17 (written by Borgan and Samuelsen from Norway), aimed at statisticians

Analysis of case-cohort design: Idea

- Main idea in analysis of case-cohort data is **weighting of observations**.

- **All cases in the cohort** are included in the case-cohort sample:
  - Each case has weight = 1 in the analysis of the case-cohort sample

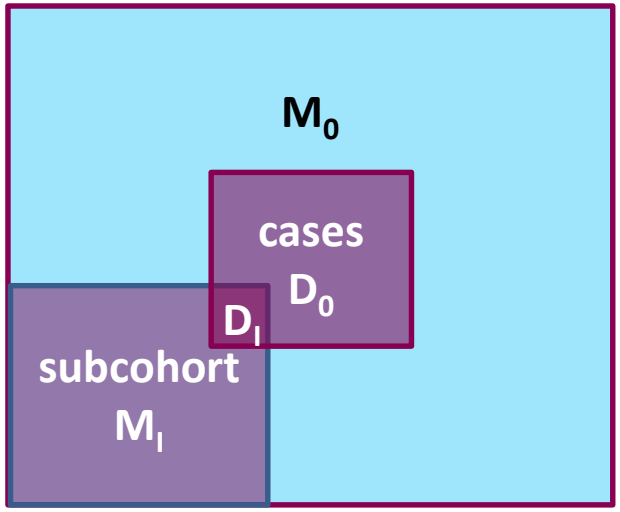
- A **sample of non-cases from the cohort** are included in the case-cohort sample:
  - Each non-case has weight w=1/p_M *(one over the sampling fraction of non-cases)*
  - All non-cases are upweighted so that each sampled non-case represents 1/p_M non-cases in the full cohort (if p_M=5% then 1/p_M=20)
  - Since subcohort is selected randomly, the upweighted case-cohort sample will be very similar to the full cohort, and representative of full cohort with respect to follow-up and exposures

- **By weighting the case-cohort data, we get inference for the full cohort!**

Analysis of case-cohort design: Sampling fractions

- We need to keep track of persons inside/outside subcohort and cases/noncases

	Outside subcohort	Inside subcohort	Total
Non-case	M_0	M_I	M
Case	D_0	D_I	D
Total	**N_0**	**N_I**	**N**



Sampling fraction: $p = \frac{N_I}{N} = \mathbf{0.05}$

Sampling fraction non-cases: $p_M = \frac{M_I}{M} \approx \mathbf{0.05} = \frac{N_I}{N}$

Sampling fraction cases: $p_D = \frac{D_0 + DI}{D} = \mathbf{1}$

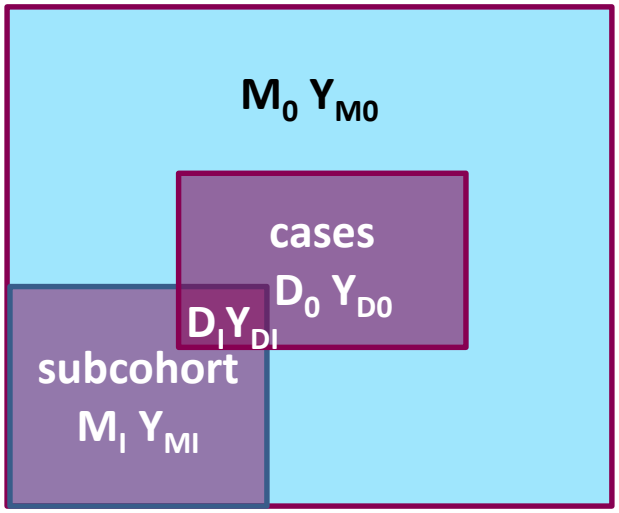
Because the full cohort is enumerated, we know M_0, M_I, D_0, D_I

However, exposure will only be known for M_I, D_0, D_I

Analysis of case-cohort design: Estimating rates (non-parametric)

Estimated by Y_{MI}/p

	Outside subcohort	Inside subcohort
Non-case	M_0	M_I
	Y_{M0}	Y_{MI}
Case	D_0	D_I
	Y_{D0}	Y_{DI}



- The Y_{M0} pyrs for non-cases outside the subcohort are known overall, but not for exposed/unexposed *(see next slide)*
- However, since we know the sampling fraction *p*, we can estimate the pyrs in the cohort

$$Y = Y_{M0} + Y_{MI} + Y_{D0} + Y_{D1} \approx Y_{MI}/p + Y_{D0} + Y_{D1}$$

- Hence, $\lambda = D/Y \approx (D_0 + D_I) / (Y_{MI}/p + Y_{D0} + Y_{D1})$

- **Estimated rate from case-cohort is a consistent estimate of the full cohort rate**

Analysis of case-cohort design: Estimating rate ratios

Exposed	Outside subcohort	Inside subcohort
Non-case	M_0	$\mathbf{M_I}$
	Y_{M0}	$\mathbf{Y_{MI}}$
Case	$\mathbf{D_0}$	$\mathbf{D_I}$
	$\mathbf{Y_{D0}}$	$\mathbf{Y_{DI}}$

Unexposed	Outside subcohort	Inside subcohort
Non-case	M_0	$\mathbf{M_I}$
	Y_{M0}	$\mathbf{Y_{MI}}$
Case	$\mathbf{D_0}$	$\mathbf{D_I}$
	$\mathbf{Y_{D0}}$	$\mathbf{Y_{DI}}$

- The exposed Y_{M0} (unexposed Y_{M0}) pyrs for non-cases outside the subcohort are unknown
- However, since we know the sampling fraction p, we can estimate

  The exposed pyrs in the cohort: $\mathbf{Y_{exp}} = Y_{M0} + \mathbf{Y_{MI}} + \mathbf{Y_{D0}} + \mathbf{Y_{D1}} \approx \mathbf{Y_{MI}}/\boldsymbol{p} + \mathbf{Y_{D0}} + \mathbf{Y_{D1}}$

  The unexposed pyrs in the cohort: $\mathbf{Y_{unexp}} = Y_{M0} + \mathbf{Y_{MI}} + \mathbf{Y_{D0}} + \mathbf{Y_{D1}} \approx \mathbf{Y_{MI}}/\boldsymbol{p} + \mathbf{Y_{D0}} + \mathbf{Y_{D1}}$
- Hence, HR= $(D_{exp}/Y_{exp}) / (D_{unexp}/Y_{unexp}) \approx D_0+D_I / (\mathbf{Y_{MI}}/\boldsymbol{p} + \mathbf{Y_{D0}} + \mathbf{Y_{D1}}) /$

  $D_0+D_I / (\mathbf{Y_{MI}}/\boldsymbol{p} + \mathbf{Y_{D0}} + \mathbf{Y_{D1}})$
- **Estimated HR from case-cohort is consistent estimate of the full cohort HR**

Case-cohort design: Weighted Cox regression (model estimation)

- To account for the under-sampling of non-cases the Cox partial likelihood must include weights, w_i

$$L(\beta)=\prod_{j=1}^{k}\frac{\exp(\beta x_j)}{\sum_{i\in R'j}\exp(\beta x_i)w_i}$$

- The risk sets represent case-cohort sample risk sets, $R'j$, i.e. subcohort plus cases outside subcohort.
- Based on theory of **inverse probability weighting** (IPW)
- A weighted likelihood is a ***pseudo-likelihood***, can be used for estimating parameters and CIs, but likelihood ratio tests are not valid (Wald tests OK)
- Need to correct standard errors (the pseudo-likelihood is upweighting the same individuals, too little variation) using robust std err (e.g. sandwich estimator)

Case-cohort design: Analysis

- The literature has focused on modifications of the partial likelihood in the Cox model. (Parametric models can also be used.)

- Design and methodology was proposed by Prentice 1986.
  - Previous work by Kupper et al (1975) and Miettinen (1982)

- Several types of weighting schemes have been proposed
  - See Kulathinal et al (2007), good overview
  - Not all weights give inference for the full cohort

- In this lecture, I focus on Borgan II weights (Borgan et al, 2000)
  - For cases: w=1
  - For non-cases: w=1/p_M *(one over the sampling fraction of non-cases)*

How to in R: Define the cohort

- Generating a case-cohort study is very easy in R.
- Ensure the cohort is defined: start and end of follow-up, risktime, event

- Example Colon cancer, localised (stage=1), cause-specific survival (status=1)

```
> colon$event <- (colon$status==1)  ## creates indicator 0/1
> localised <- subset(colon, stage == 1)
```

- Follow-up: **`surv_mm`**
- Event indicator: **`event`**

How to in R: Create the case-cohort sample

```
> set.seed(42)                                   ## make sampling reproducible

## sample subcohort
> localised$u <- runif(6274, 0,1)                ## assign random number to all obs
> localised$subcoh[localised$u<=0.05] <- 1       ## generate dummy subcohort 5%
> localised$subcoh[localised$u>0.05] <- 0
> table (localised$event, localised$subcoh, useNA="always")

           0    1 <NA>
  FALSE 4338  202    0
  TRUE  1660   74    0
  <NA>     0    0    0
```

event	subcoh 0	1	Total
0	4338	202	4,540
1	1660	74	1,734
Total	5998	276	6,274

Full cohort: n= 6274

Case-cohort: n= 1936 (i.e. 1660+202+74)

Sampling fraction non-cases:

$$p_M = \frac{202}{4540} = 0.04449$$

Sampling fraction, total:

$$p = \frac{276}{6274} = 0.04399$$

How to in R: Generate weights

```
## generate Borgan II weights
> localised$wt[localised$event==1] <- 1  ## wt=1 if case==1
> localised$wt[localised$event==0 & localised$subcoh==1] <- 1/(244/4540)
                                        ## wt=1/samplfrac if case=0, subcoh=1

> table (localised$wt, useNA="always")
                1      22.4752475247525              <NA>
             1734                   202              4338
```

How to in R: Weighted models

```
> summary(coxph(Surv(surv_mm,event)~sex + factor(agegrp) + year8594,
                  data=localised, ties="breslow", weights = wt, robust = TRUE ))

   n= 1936, number of events= 1734
   (4338 observations deleted due to missingness)

                    coef exp(coef) se(coef) robust se      z Pr(>|z|)
sex              -0.14574   0.86438  0.04937   0.14028 -1.039   0.2988
factor(agegrp)1 -0.16520   0.84773  0.13937   0.34531 -0.478   0.6324
factor(agegrp)2  0.26276   1.30051  0.12583   0.31568  0.832   0.4052
factor(agegrp)3  0.80080   2.22731  0.12590   0.32175  2.489   0.0128 *
year8594         -0.17738   0.83746  0.04964   0.13659 -1.299   0.1941
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The estimates from case-coh (coef(sex) -0.14574) is a bit different to the full cohort (coef(sex) -0.08871)
- Standard errors are higher in case-coh (se(sex): 0.14028) compared to the cohort (se(sex): 0.04937).

How to in Stata: Define the cohort

- Generating a case-cohort study is very easy in Stata.

- Start by stsetting the data, and generating a case variable based on the event indicator from stset (_d).

```
. stset surv_mm if stage==1, failure(status==1) scale(12) id(id)

. gen case=_d  NOTE: IMPORTANT! Define case based on _d, which
                    accounts for censoring.
```

How to in Stata: Create the case-cohort sample

```
. set seed 339487732  // makes sampling reproducible
. gen u = runiform()   // assign random number to all obs
. gen subcoh = u < 0.05 // generate dummy subcohort
. tab case subcoh
```

case	subcoh 0	1	Total
0	4,335	205	4,540
1	1,652	82	1,734
Total	5,987	287	6,274

Full cohort: n= 6274

Case-cohort: n= 1939 (i.e. 205+1652+82)

Sampling fraction non-cases:

$$p_M = \frac{205}{4540} = 0.04515$$

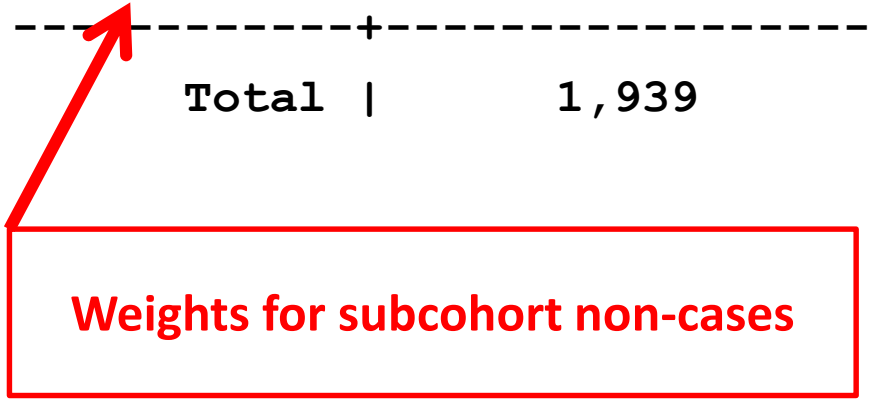
Sampling fraction, total:

$$p = \frac{287}{6274} = 0.04574$$

How to in Stata: Generate weights

```
// Generate Borgan II weights
. gen wt = 1 if case==1
. replace wt = 1 / (205/4540) if case==0 & subcoh==1
. tab wt

        wt |      Freq.     Percent        Cum.
-----------+-----------------------------------
         1 |      1,734       89.43       89.43
  22.14634 |        205       10.57      100.00
-----------+-----------------------------------
     Total |      1,939      100.00
```



How to in Stata: Weighted models

```
. /* STSET using pweights option*/
. stset surv_mm if stage==1 [pw=wt], failure(status==1) scale(12) id(id)

. /* Cox model for case-cohort - Borgan II*/
. stcox sex i.agegrp year8594, vce(robust)
```

_t	Haz. Ratio	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex	.952777	.1304472	-0.35	0.724	.7285361	1.246039
agegrp						
45-59	1.064393	.3438639	0.19	0.847	.5650824	2.004897
60-74	1.899299	.5604331	2.17	0.030	1.065188	3.386574
75+	2.28059	.6781713	2.77	0.006	1.273293	4.084757
year8594	.8036375	.1071236	-1.64	0.101	.6188657	1.043576

Comparison full cohort, NCC, case-cohort (R)

Log(HR) ***Stderr***	**Full cohort** **Cox**	**NCC (1:1)** **Conditional log reg**	**Case-cohort 5%** **Weighted Cox**
Sex	-0.08871 0.04937	-0.06357 0.07093	-0.14574 0.14028
Age 45-59	-0.05217 0.13845	-0.12520 0.18728	-0.16520 0.34531
Age 60-74	0.29155 0.12573	0.15132 0.17029	0.26276 0.31568
Age 75+	0.81025 0.12607	0.75649 0.17603	0.80080 0.32175
Year8594	-0.28121 0.04937	-0.33093 0.07176	-0.17738 0.13659
N total	6274	3468	1936
N cases	1734	1734	1734
N non-cases	4540	1734	202

Comparison full cohort, NCC, case-cohort (Stata)

HR ***stderr***	**Full cohort** **Cox**	**NCC (1:1)** **Conditional log reg**	**Case-cohort 5%** **Weighted Cox**
Sex	0.9151101 0.0451776	0.9058728 0.063661	0.952777 0.1304472
Age 45-59	0.9491689 0.1314101	0.927094 0.168337	1.064393 0.3438639
Age 60-74	1.338501 0.1682956	1.276786 0.2123023	1.899299 0.5604331
Age 75+	2.24848 0.2834768	2.268003 0.3845136	2.28059 0.6781713
Year8594	0.7548672 0.0372669	0.7763301 0.055581	0.8036375 0.1071236
N total	6274	3468	1939
N cases	1734	1734	1734
N non-cases	4540	1734	205

Comparison full cohort, NCC, case-cohort

- Point estimates of hazard ratios should be similar for all three approaches. Sampling variation may cause the HRs to differ from the full cohort.

- The standard errors should be higher in NCC and case-cohort designs, compared to full cohort, since we are including fewer observations. But the additional error is very small in comparison to the gain in dataset reduction.

- In the full cohort, there is approx 2.6 non-cases per case (1734:4540)
- In the NCC, there is 1 non-case per case
- In the case-coh, there is approx 0.12 non-case per case (1734:202)

- This affects the standard errors.
- If we instead sample 25% subcohort (approx 0.64 non-cases per case), the results are quite similar for NCC and case-cohort.

Comparison full cohort, NCC, case-cohort (R)

Log(HR) *Stderr*	Full cohort Cox	NCC (1:1) Conditional log reg	Case-cohort 25% Weighted Cox
Sex	-0.08871 0.04937	-0.06357 0.07093	-0.07999 0.08209
Age 45-59	-0.05217 0.13845	-0.12520 0.18728	0.08561 0.20382
Age 60-74	0.29155 0.12573	0.15132 0.17029	0.43927 0.18593
Age 75+	0.81025 0.12607	0.75649 0.17603	0.96799 0.18938
Year8594	-0.28121 0.04937	-0.33093 0.07176	-0.40536 0.08210
N total	6274	3468	2837
N cases	1734	1734	1734
N non-cases	4540	1734	1103

Comparison full cohort, NCC, case-cohort (Stata)

HR ***stderr***	**Full cohort** **Cox**	**NCC (1:1)** **Conditional log reg**	**Case-cohort 25%** **Weighted Cox**
Sex	0.9151101 0.0451776	0.9058728 0.063661	0.915073 0.0648774
Age 45-59	0.9491689 0.1314101	0.927094 0.168337	0.9058277 0.1690948
Age 60-74	1.338501 0.1682956	1.276786 0.2123023	1.27769 0.218906
Age 75+	2.24848 0.2834768	2.268003 0.3845136	2.123809 0.3693696
Year8594	0.7548672 0.0372669	0.7763301 0.055581	0.7181138 0.0496757
N total	6274	3468	2838
N cases	1734	1734	1734
N non-cases	4540	1734	1104

Comparison full cohort, NCC, case-cohort

- The statistical efficiency of NCC and case-cohort studies are similar, given the same number of non-cases per case.

- Hence, it is important to choose a sampling fraction that give between 1 to 5 non-cases per case. (Not much gain in having more than 5 controls per case.)

Other measures of interest, e.g. rates

- An advantage of the case-cohort design is the possibility to obtain estimates of the rates (not just rate ratios) in the underlying cohort.

- The amount of risktime in the underlying cohort can be estimated via the sampling fraction.

- Hence, the rate can be estimated from a model (preferably parametric model, such as Poisson regression) or crudely by summing cases and risktime.

Some examples of case-cohort studies

- Search on Pubmed: “case-cohort”
- Motivation for all these studies to use case-cohort: Expensive exposures!

	Exposure	Outcome	
Geybels et al; 2014 J Natl Cancer Inst	• Genes interacting with levels of proteins; • Genotyping on subcohort+cases	• Advanced prostate cancer • Only one outcome!	Dutch cohort study Case-cohort design
Luft et al; 2015 Diabetic Medicine	• Levels of carboxymethyl lysine (CML)/biomarker; • Lab analysis on subcohort+cases	• Diabetes • Only one outcome!	US cohort study **Stratified case-cohort** (stratified on race: 50% African American, 50% white)
Karvanen et al, 2009; MORGAM project Genet Epidemiol.	• Genes; • Genotyping on subcohort+cases	• CHD, Stroke, total mortality • **Many outcomes**	Multi-national cohort collaboration Case-cohort design

- Johansson et al, Breast Cancer Res Treat (2015): Case-cohort design in a register-based study to improve computational efficiency after multiple timescale splitting.

In summary: Nested Case-Control vs. Case-Cohort

Nested Case-Control (NCC)	Case-Cohort
Matched on time, only one outcome	No time matching, more than one outcome possible
Closed or Open (delayed entry) cohorts; riskset sampling valid	Closed cohorts (sampling at entry), or open cohorts; sampling of follow-up times
Simple to analyse, but absolute risks/rates are complicated to obtain	Semi-complicated to analyse, but absolute risks/rates are easy to obtain
Matched on one timescale (no main effect estimable, but interactions are estimable); multiple timescales possible (but often matched on other timescales)	Multiple timescales (both main effects and interactions estimable); flexibility to change and choose timescales in analysis
HR can be estimated	HR and hazards, hazard differences, cumulative risk; information about underlying cohort/population at risk is maintained via the sampling fraction
More common in literature	Less common in literature

References

- **Cox DR (1972)** Regression models and life-tables. J Royal Stat Soc 1972
- **Thomas (1977)**
- **Prentice, Breslow (1978)** Retrospective studies and failure time models. Biometrika 1978
- **Oakes (1981)** Survival times: Aspects of partial likelihood. Int Stat Rev 1981
- **Goldstein, Langholz (1992)** Assymptotic theory for nested case-control sampling in the Cox regression model
- **Greenland, Thomas (1982)** On the need for the rare disease assumption in case-control studies. Am J Epi 1982
- **Klein JP, van Houwelingen HC, Ibrahim JG, Scheike TH (2013**). Handbook of survival analysis. Chapman and Hall/CRC Press, Boca Raton. (Chpt 17 by Borgan O, Samuelsen SO)
- **Prentice RL (1986);** A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika, 73:1-11. 1986.*
- **Kulathinal, Karvanen, Saarela, Kuulasmaa (2007);** Case-cohort design in practice – experiences from the MORGAM project. *Epidemiol Perspect Innov, 2007*.
- **Moger, Borgan, Pawitan (2008)**; Case–cohort methods for survival data on families from routine registers. *Statist in Med, 27(7): 1062-1074. 2008.*
- **Langholz, Borgan (1997)**; Estimation of absolute risk from nested case-control data. *Biometrics, 1997.*
- **Samuelsen;** *teaching notes from 2005* - http://folk.uio.no/osamuels/casecohort4.pdf
- **Borgan, Samuelsen (2003):** A review of cohort sampling designs for Cox's regression model: Potentials in epidemiology. *Norsk Epidemiologi, 13:239-248. 2003*
- **Cologne et al (2012);** Conventional case–cohort design and analysis for studies of interaction. *International Journal of Epidemiology 2012;1–13*

References

Examples of epi studies which have used the case-cohort design:

- **Karvanen et al (2009);** The impact of newly identified loci on coronary heart disease, stroke and total mortality in the MORGAM prospective cohorts. *Genet Epidemiol, 2009*.
- **Luft et al (2015);** Carboxymethyl lysine, and advanced glycation end product, and incident diabetes: a case-cohort analysis of the ARIC study. *Diabetic Medicine 2015*
- **Geybels et al (2014);** Selenoprotein gene variants, toenail selenium levels, and risk for advanced prostate cancer. *JNCI, 2014*

- **Johansson et al (2015);** Family history and risk of pregnancy-associated breast cancer, Br Ca Res Treat 2015