Survival analysis in other study designs:

Nested case-control
Case-cohort
Matched cohort

Anna Johansson

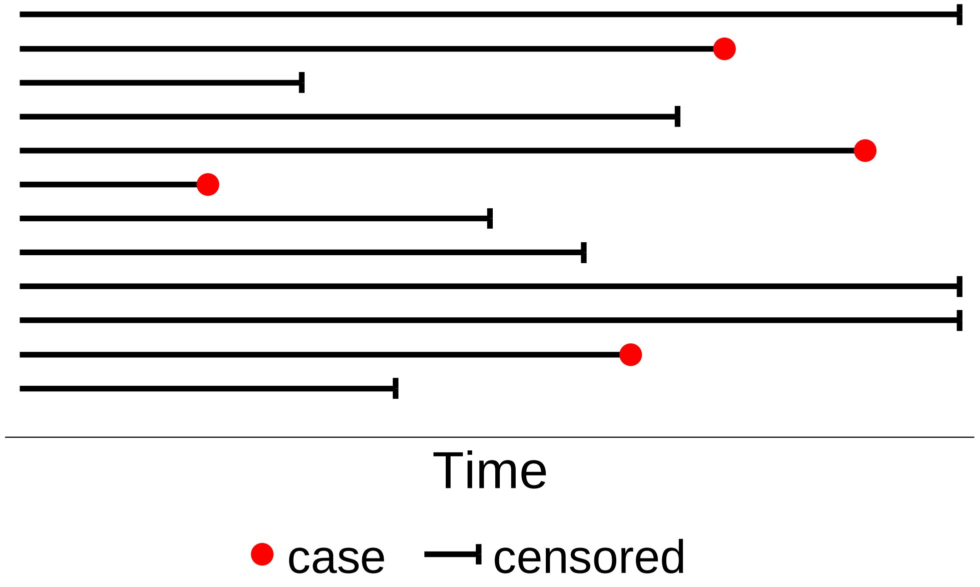
Biostat 3

(version 2025-01-10, v1)

Cohort study

- A cohort study is characterised by a **group of individuals, which are followed** for a specific outcome over time.
- Cohort members are assumed to be **free of the outcome** (”disease-free”) at start of follow-up.
- Cohort members are followed until they have the outcome or they are censored (no longer under follow-up).
- From the cohort, we can estimate event rates (hazard rates) and relative rates (hazard ratios).

- We can fit a survival model, e.g. Poisson regression or Cox regression, to the data and obtain estimates of the hazard rates (Poisson) and the hazard ratios (Poisson, Cox).



Cohort study: Cox regression with maximum likelihood estimation

- In Cox regression, the hazard rate is modelled as: **$\lambda(t|X)= \lambda_0(t) \times \exp(\beta X)$**
- 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.

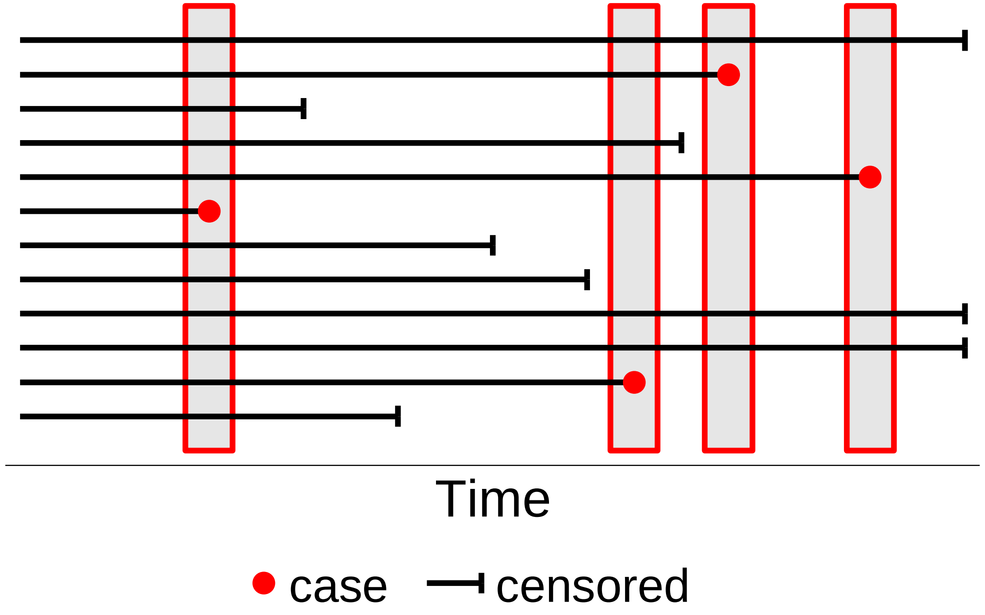
Cohort study: Cox regression, risk sets and likelihood

- 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 **λ(t|X)= λ$_0$(t) × exp(βX)** and the **λ$_0$** cancel out, we can write this as:

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



Cohort study: Cox regression, risk sets and likelihood

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

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

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

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

Cohort study: Cox regression 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.

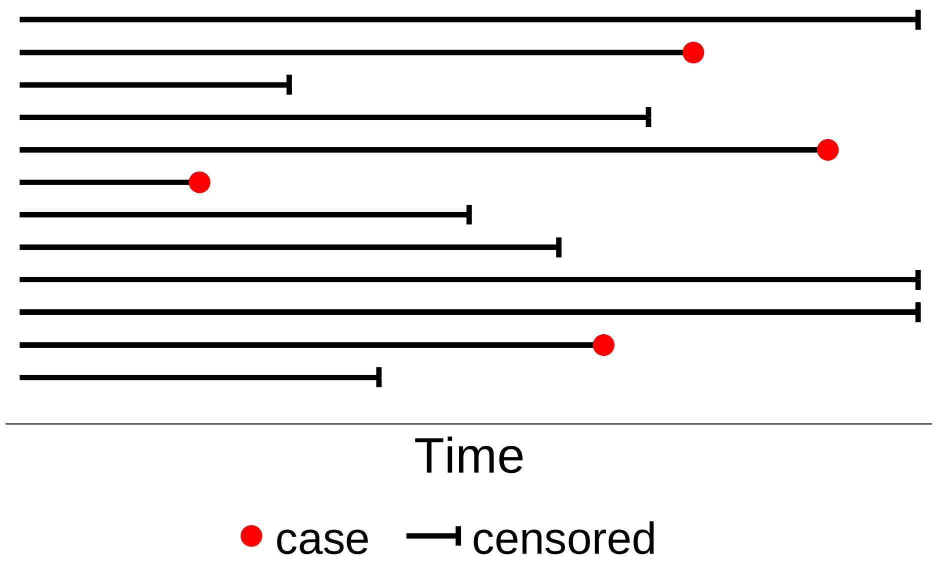
Alternative designs: Case-control studies

- In some situations, we may be unable to use a full cohort.
- E.g., we may want to collect expensive information on exposures or confounders (e.g. genotyping, questionnaires), which is not feasible on the full cohort due to cost and time.
- E.g., we may want to reduce the analytical sample (data sizes) for computational efficiency (complex modelling).
- Then we may use a sampling design, which reduces the study size.
- Two such designs are the **nested case-control design (NCC)** and the **case-cohort design**.

- Both these designs select **all cases** (events in the cohort) and a **sample of controls**, i.e. sampling on outcome status.
- Both the NCC and the case-cohort designs capture the information from the **underlying full cohort**. The results (inference) are interpretable for the full cohort from which they were sampled.
- With appropriate sampling and analysis, the **odds ratio from a NCC design** estimates the **hazard ratio in the cohort**.
- With appropriate sampling and analysis, **hazard rates and hazard ratios from the case-cohort design** estimates the **hazard rates and hazard ratios in the cohort**.

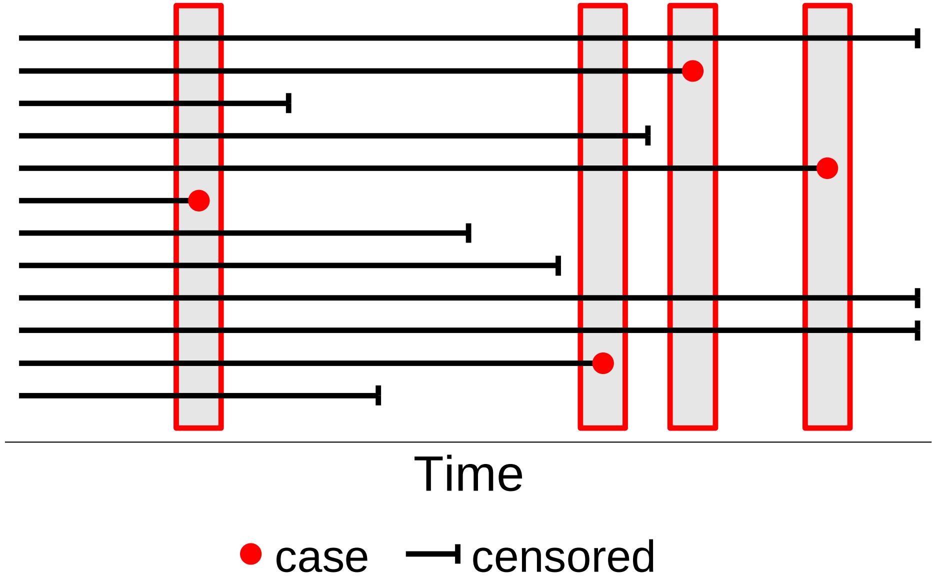
Nested case-control: Sampling

- We start with a cohort study.



Nested case-control: Sampling

- We start with a cohort study.
- Create risk sets around the events.

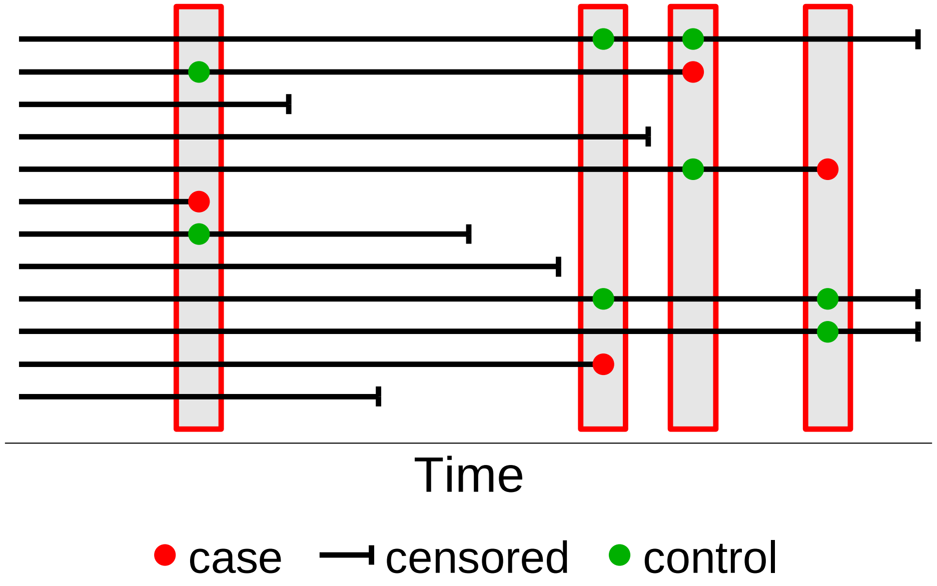


Nested case-control: Sampling

- We start with a cohort study.
- Create **risk sets** around the events.

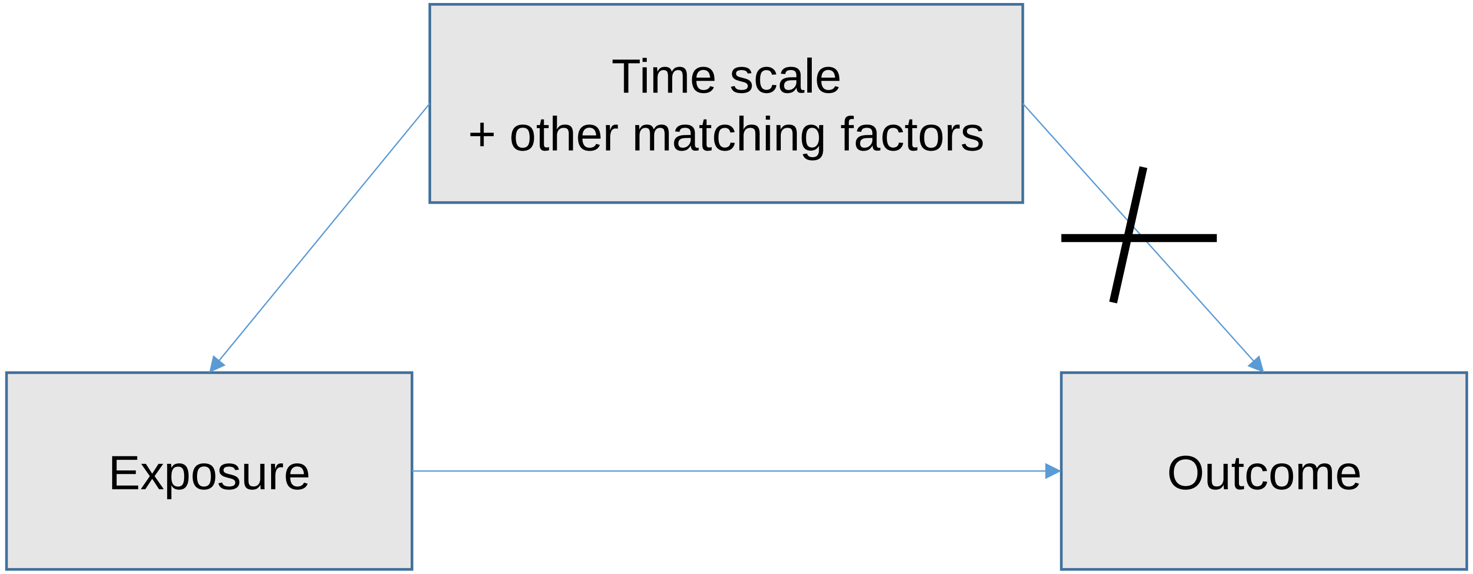
Sampling of controls:

- Select a **fixed number of controls** randomly from each risk set.
- Usually 1-5 controls per case (>5 controls only improves statistical power minorly).
- Persons can be selected as controls to several cases.
- A person selected as a control may later become a case.
- The case-control sample consists of all cases and their matched controls.
- Controls are tightly **time-matched to cases** due to risk sets. Controls can typically only be used for one type of outcome.
- Matching on time (time scale) is sensible if the **time scale is a confounder** and **no interest in estimating the effect of time**.



Nested case-control: Analysis

- Matching in NCC studies usually also involves **matching on other factors**, e.g. age, sex or region of diagnosis.
- The controls are selected with the **same distribution of matching factors** as the cases, e.g. each control is selected with same age, sex and region as its case.
- Matching introduces a **selection bias**, which is then **removed with a conditional analysis**.
- Therefore matching (<u>sampling</u> and <u>conditional analysis</u>) removes the association between the matching factors and the outcome.



Nested case-control: Analysis

- Originally proposed by Thomas (1977), also developed by Prentice and Breslow (1978), Oakes (1981), Goldstein and Langholz (1992).

- The analysis is conditioning on risk set (and other matching strata).
- The **NCC partial likelihood** is very similar to the Cox partial likelihood for the full cohort.

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

- $\tilde{R}j$ is the case-control sampled risk set, rather than the full risk set Rj.
- Interestingly: The **NCC partial likelihood** is identical to 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 risk set (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: 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: 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: 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: Stata

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

- It may be non-intuitive that a person can be both a case and a control in the same study.
- However, by looking at the likelihood for the Cox model (as created from risksets) – and by considering that a NCC study as being derived from a cohort study – it is obvious that it must be this way.
  - A person will contribute to several risksets as long as he/she is still under observation
  - A person can therefor be a control in several risksets, and also if later a case

- *Sampling among ”never cases” – why is it incorrect: Sampling on ”never cases” means that the sampling at time t is conditioning on the future (this will not give incidence). The risk set sampling ensures the estimates are not depending on the future (which may introduce survival bias).*

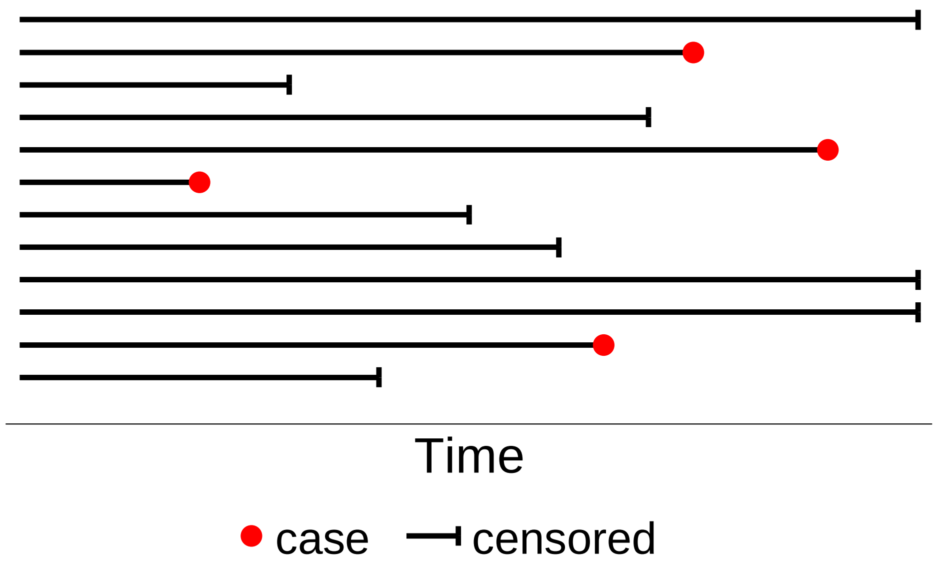
Nested case-control: Limitations

- 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 risk set then controls can be re-used.*

- Limitation 2:
  - We can **only estimate hazard ratios** from a NCC design.
  - We **cannot estimate hazard 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 risk set).
  - *Not entirely true. If we know the size of risk sets and sampling fractions in each risk set, then it is possible to estimate rates (Langholz, Borgan 1997 and others). Not trivial!*

Case-cohort: Sampling

- We start with a cohort study.



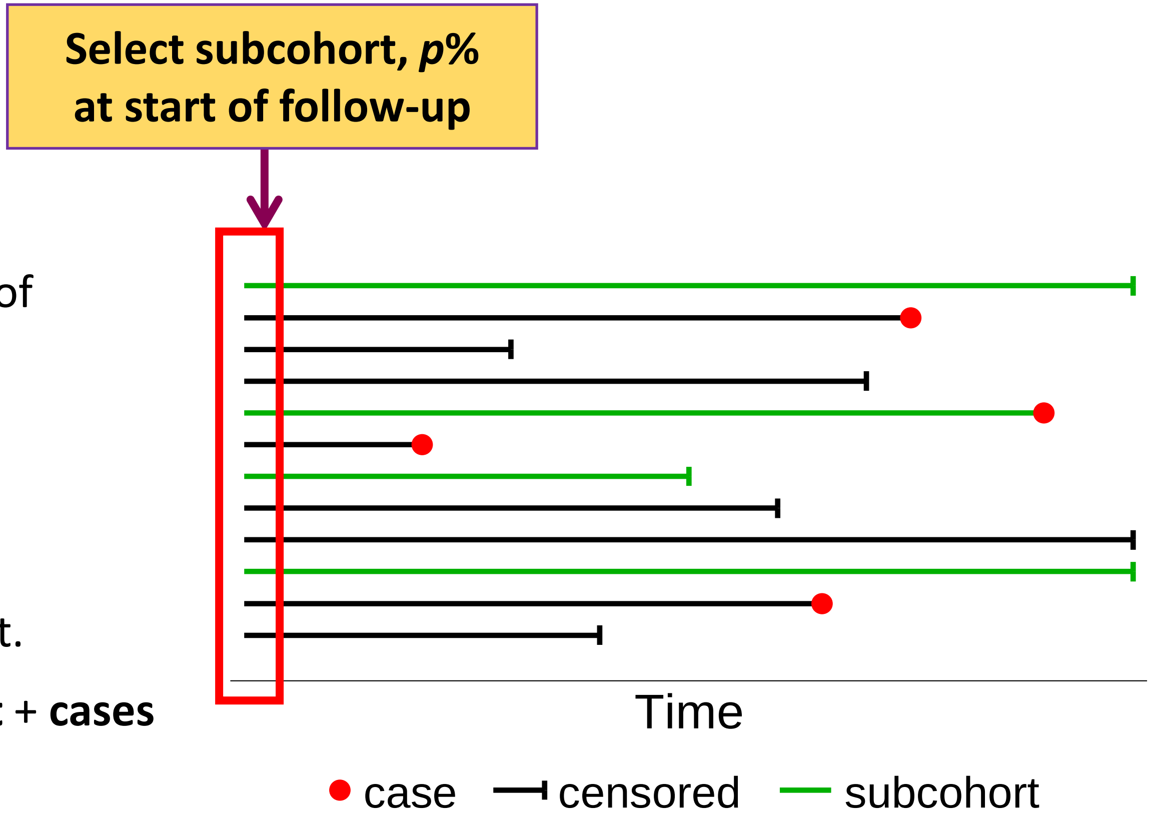
Case-cohort: Sampling

- We start with a cohort study.

Sampling of controls:

- Select a random sample of the cohort, of p%, at start of follow-up. This is the **subcohort**.
- Although we say ”start of follow-up”, we include all persontime, i.e. we select the ”lines” of follow-up.
- The subcohort will include some cases.
- Also include all cases that occur outside the subcohort.
- The final **case-cohort sample** consist of the **subcohort** + **cases outside** the subcohort.

- The subcohort is **<u>not</u> time-matched** to cases. This means that it can be used for many types of outcomes.
- The case-cohort sampling also works with delayed entry, i.e. start of follow-up may differ between persons.



Select subcohort, *p%* at start of follow-up

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

Case-cohort: Limitations

- 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

Case-cohort: Analysis

- 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 w = 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)*
  - This means that all non-cases are upweighted so that each sampled non-case represents 1/p_M non-cases in the full cohort.
  - E.g. if p_M=5% then 1/p_M=20 meaning that each sampled non-case represents 20 non-cases in the full cohort.

- 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
- Hence, by weighting the case-cohort data, we get **inference for the full cohort**.

Case-cohort: Analysis

- 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 **sampled risk sets**, $R'j$, i.e. subcohort plus cases outside subcohort.
- This approach is based on theory of inverse probability weighting (IPW).
- A weighted likelihood is a ***pseudo-likelihood***, which can be used for estimating parameters and CIs, but likelihood ratio tests are not valid (Wald tests OK).

- Additionally, due to that the same subcohort is used over time in several risks sets, the **risk sets are correlated**.
- This means that we need to correct the standard errors (the pseudo-likelihood is upweighting the same individuals, too little variation) using ***robust standard errors*** (e.g. sandwich estimator).

Case-cohort: Stata

- 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.

. 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

           |        subcoh
      case |         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:

$$\boldsymbol{p = \frac{205}{4540} = 0.04515}$$

Sampling fraction, total:

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

Case-cohort: Stata

```
// Generate weights (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
```

- Weights for non-cases are 22.14634, meaning that each sampled non-case represents just over 22 non-cases in the full cohort.

Case-cohort: Stata

- The case-cohort sample must be stset using weights, and then the stcox command will automatically use the weights.
- Standard errors can be corrected by using the vce(robust) option.

```
. /* 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)

------------------------------------------------------------------------------
             |               Robust
          _t | Haz. Ratio   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, Nested case-control and Case-cohort (Stata)

	Full cohort Cox	NCC (1:1) Cond. log reg	Case-cohort 5% Weighted Cox	Case-cohort 25% Weighted Cox
Sex: HR Sex: std err	0.9151101 0.0451776	0.9058728 0.063661	0.952777 0.1304472	0.915073 0.0648774
N total	6274	3468	1939	2838
N cases	1734	1734	1734	1734
N non-cases	4540	1734	205	1104

- **Point estimates of hazard ratio** 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-cohort, there is approx 0.12 non-case per case (1734:205)
- If we instead sample 25% subcohort (approx 0.64 non-cases per case), the results are similar for NCC and case-cohort.
- Hence, given the **same number of non-cases per case**, the statistical efficiency of NCC and case-cohort are similar.

Comparison 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 in both.	Closed cohorts (sampling at entry), or open cohorts; sampling of follow-up times valid in both.
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.
HRs can be estimated.	HR and hazard rates, hazard differences, cumulative risk can be estimated, since information about underlying cohort/population at risk is maintained via the sampling fraction.
More common in literature.	Less common in literature.

References: Nested case-control and Case-cohort

Cox DR (1972). Regression models and life-tables. *J Royal Stat Soc 1972*

Thomas D (1977). Addendum to ‘Methods of cohort analysis: Appraisal by application to asbestos mining’ by Liddell FDK, McDonald JC, Thomas DC. *Journal of the Royal Statistical Society. 1977;A 140:469–91.*

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.*

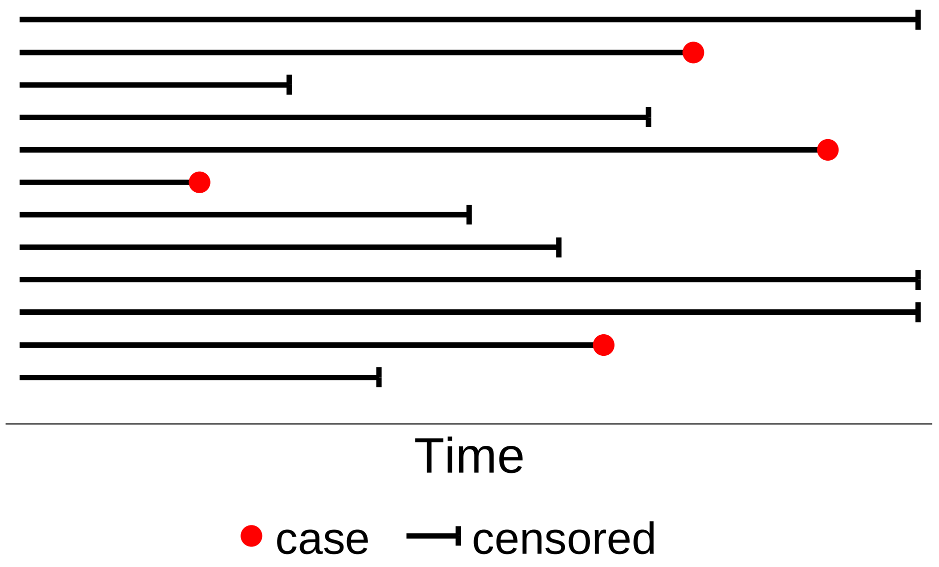
Samuelsen (2005). *Teaching notes from 2005.*

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*

Matched cohort: Sampling

- We start with a cohort study.

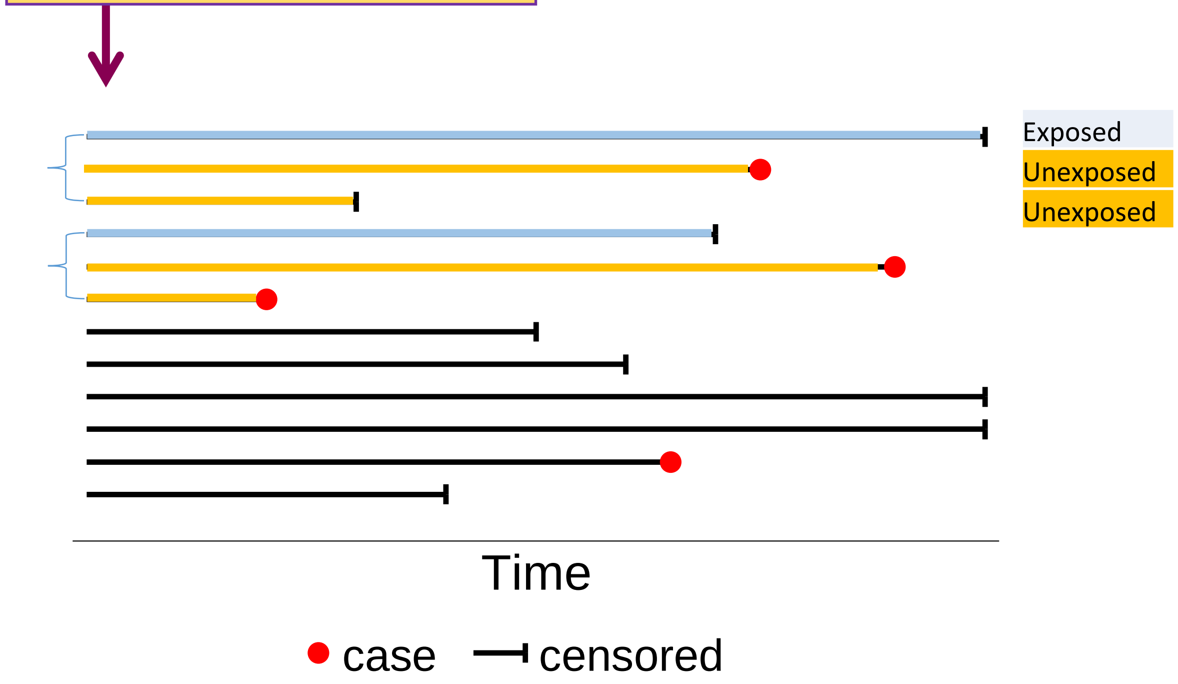


Matched cohort: Sampling

- We start with a cohort study.
- Matching in a cohort study means that **exposed** persons are matched to **unexposed** at time of exposure.
- Time of exposure (matching) is also start of follow-up.

- Other matching factors (e.g. age, sex) means that the **unexposed** are selected to have the **same distribution** of those matching factors **as the exposed**.
- Select a **fixed number of unexposed** randomly for each exposed. Usually 1 to 5 unexposed per exposed, but may be useful to select >5 unexposed, if the outcome is rare.
- By randomly selecting the unexposed with similar distributions on matching factors as the exposed, the **association between the exposure and matching factors is eliminated** (**at start of follow-up**), and the confounding is thus removed.

Select unexposed at start of follow-up

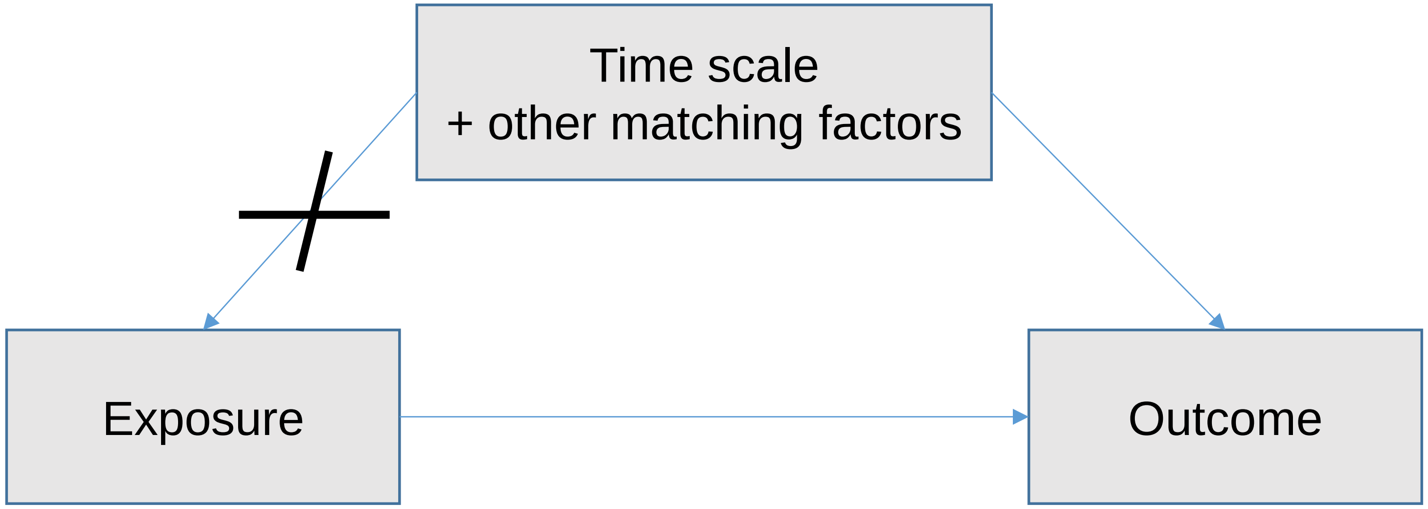


Matching is on timescale. If date of diagnosis, then matched on calendar period. If age at diagnosis, then matched on attained age.

Delayed entry is possible. I.e. matched sets could start follow-up at different times.

Matched cohort: Confounding eliminated

- **Elimination of confounding** is the main reason for matching in cohort studies.
- The **association between matching factors and exposure <u>at start of follow-up </u>is eliminated** and the matching factors do not act as confounders for the exposure-outcome association.
- Unlike matched case-control studies (which introduces a selection bias), matching in cohort studies will remove confounding at the start of follow-up (no bias introduced in the design).
- In a cohort study with no distortion* in the matching balance during follow-up, **no additional action in the analysis is required to control for confounding** of the crude point estimate, because matching prevents an association between exposure and the matching factor.



* No competing risks, no informative loss to follow-up (censoring) and no confounding during followup.

Matched cohort: Purpose of matching

Several reasons for matching in cohort studies:

- Conceptual simplicity in **controlling for confounding** (at start of follow-up) – crude comparisons between exposed and unexposed are controlled for confounding by matching factors, i.e. we compare like-with-like in Table 1, easy to communicate.
- **Interest in causal effect** among exposed – e.g. then matching unexposed to exposed is sensible.
- **Cost-efficiency** in exposure collection – e.g. if additional covariates or outcomes are costly to obtain, then matching will increase cost-efficiency.
- **Computational efficiency** – e.g. by reducing the analytical sample, more complex statistical models.

- Statistical efficiency is not a reason for matching – no general rule for when matched cohort studies are more statistically efficient than the full cohort analysis.
- Impact on efficiency can vary depending on the effect measure. (Greenland & Morgenstern, 1990)

Matched cohort: Adjusting for matching is often needed

- Matching in cohort studies **controls for confounding at start of follow-up**.
- Importantly: **If there are factors that distorts the matching balance <u>during follow-up</u>**, i.e. distorts the ”exposure ↔ matching factors” association or the ”outcome ↔ matching factors” association, such as

    (1) competing risks,

    (2) loss to follow-up (censoring),

    (3) confounding during follow-up,

then the analysis of a matched cohort study **<u>must be adjusted</u> for factors associated with those** (1)-(3) in a similar way as a normal cohort study must account for (1)-(3).

- In particular, if there are additional confounders other than matching factors, either at baseline or during follow-up, that need to be adjusted for in the analysis, then matching cannot be ignored in the analysis
- **Hence, in practice, as we tend to have competing risks, censoring and/or additional confounding during follow-up, the standard analysis of a matched cohort study is to include adjustment for the matching factors.**

(Rothman KJ, Greenland S, Lash TL.)

(Sjölander A, Greenland S. 2013)

Matched cohort: Adjusting for matching

- To adjust for matching in the analysis, each matched set is treated as a distinct stratum in a stratified or conditional analysis, e.g. stratified Cox regression or Cox regression with matching strata added as covariates.
- **Stratified Cox regression** allows each matched set to have its own baseline. This works even if each matched set is small, since the baseline is not estimated.

```
. stcox i.exposure i.educ i.year , strata(stratum_id)
```

- **Cox regression with matching strata added as covariates** (one dummy variable for each matched set) is only sensible if the strata are few, as each strata will require one parameter each.

```
. stcox i.exposure i.educ i.year i.stratum_id
```

Matched cohort: Combine similar matching strata improves efficiency

Matching strata with similar values:

- If two or more matched sets have **identical values for all matching factors**, then matched sets can (and should) be **combined into a single stratum** in the analysis, which will improve the efficiency.

Missing data:

- Note: If some subjects have incomplete data on some covariates, and thus will be excluded in a complete-case analysis, then **combining matching** strata will also ensure that those sets which may lose an exposed/unexposed due to missingness are still included, as they can borrow exposed/unexposed from other strata.
- The slight shift in matching balance that such missingness may introduce is usually not a major flaw to the efficiency of confounder adjustment. If missingness is extensive, this may however be an issue.

Matched cohort: Interpretation of effects

- Matching on exposure status will change the risks in matched unexposed group to what would have occurred among the unexposed population if they were similar to the exposed population (assuming no unmeasued confounding).
- E.g. if the exposed group is mainly older persons, then the matched unexposed group will also be an older population but without the exposure.

- The **crude exposed vs. unexposed** comparison is no longer valid for the whole population.
- Instead, the **crude exposed vs. unexposed comparison** is valid for an underlying population that is similar to the exposed population with respect to the matching factors, e.g. for an older population.

- Sjölander et al (2012) has compared this to standardization of effects where both groups are standardized to the confounder distributions among the exposed.

”Under exposure driven matching, the induced marginal distribution of Z equals the source population distribution of Z among the exposed, that is, Pr(Z) = Pr(Z|X = 1).”*

Matched cohort: Sampling of matched cohort studies in practice

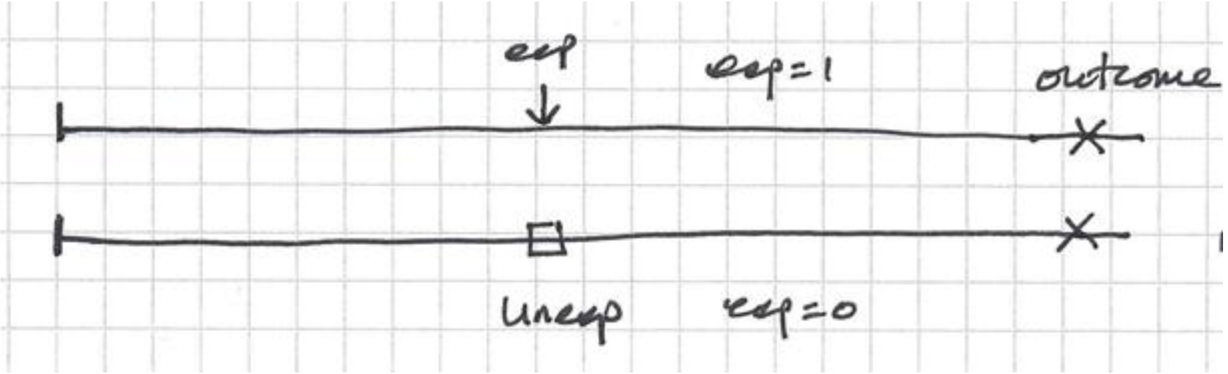
- Often expensive to do matched cohort studies, hence few international matched cohort studies – major exception is for register-based cohorts, where matching is usually cheap and straight-forward.

- Common matching factors are:
  - Age (birthyear), sex, calendar year, geographical region, hospital.
  - If matching on birthdate and date at start of follow-up, then age is automatically matched for.
  - Siblings and twins, and similar family designs – here matching strata reflect a ”clustering”.

- To consider when sampling a matched cohort:
  - Sampling the unexposed to exposed – **at what time point**? – ANSWER: **at time of exposure**!
  - Sampling **with** or **without replacement**? – ANSWER: **with** replacement!
  - *If sampling with replacement, should you censor at the time a person is re-selected to another matched set?*
  - *Time-varying exposures – at what time point should they be sampled?*

Matched cohort: When to sample the exposed and unexposed – EXTRA SLIDE

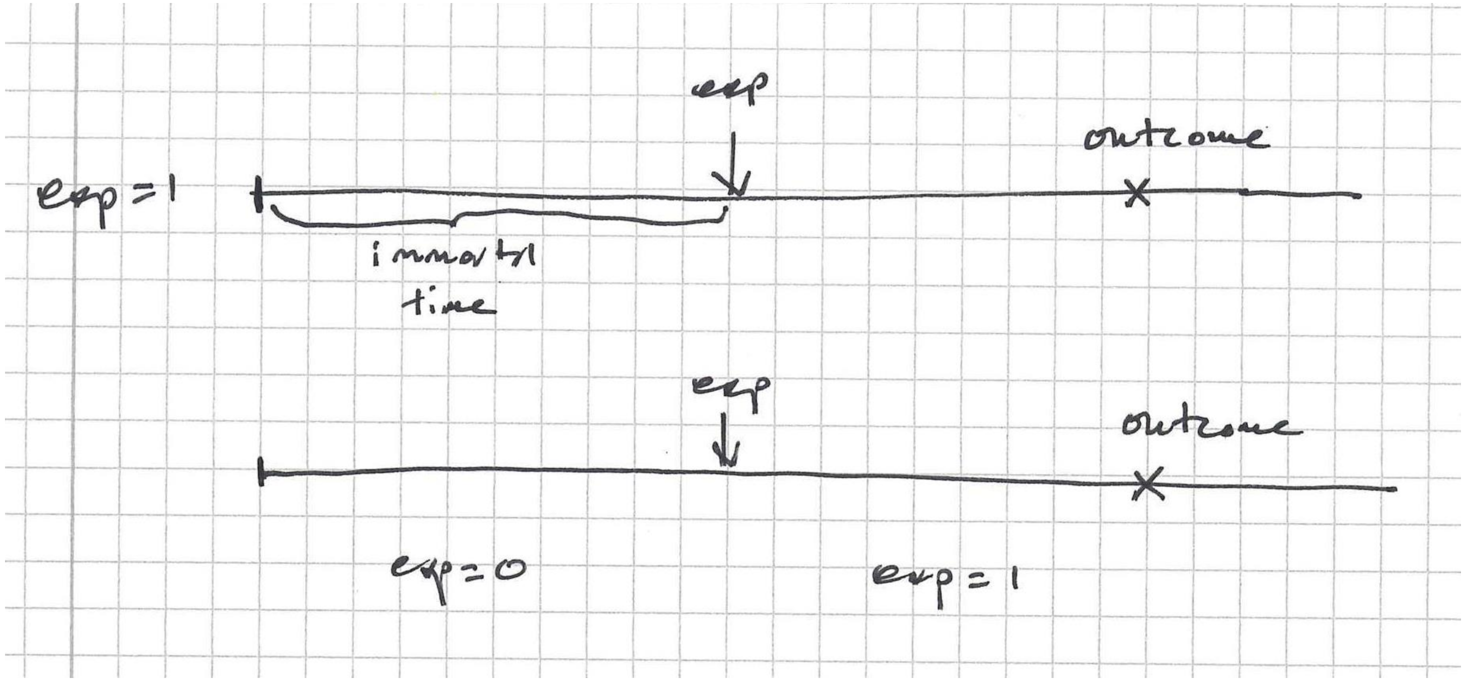
- Matching an unexposed (with similar values on matching factors) at the time of exposure (start of follow-up) seems sensible.



- A general rule is to avoid conditioning (or matching) on the future.
- An example of matching on the future is if we match exposed to ”never exposed”.
  - To classify a person as ”never exposed” one must look into the future and assess the full follow-up/exposure history.
  - Persons with shorter follow-up more likely to be ”never exposed”, have less time (chance) to become exposed.
  - Persons with shorter follow-up are also more likely to be events (i.e. not followed to study end/censoring).
  - Will lead to a selection bias (in literature, this is sometimes called *immortal time bias*).
  - *Immortal time* is a time period when a person could not have the outcome, here we are rather looking at a time period when a person could not become exposed and contribute with exposed risk time.

Immortal time bias – EXTRA SLIDE

- Immortal time bias is a period of time when a person cannot die (or have the event) since we know he/she had exposure (or a factor) happening later
  - For a person where we know that the exposure occurred at a given timepoint, that person could not have died prior to the exposure
  - All the time prior to exposure is thus immortal
  - If we classify all the time in the exposed person as ”exposed person-time”, then that mortality rate will be underestimated since the person could only have the event death after he/she was exposed – this creates immortal time bias in the mortality rate
  - A solution to immortal time bias is to split the follow-up and classify the first part as ”unexposed” and the second part as ”exposed”



Matched cohort: Sampling with or without replacement – EXTRA SLIDE

- In a case-control design (riskset sampling), sampling should be done with replacement and a person who later becomes a case is allowed to be selected as a control in previous risk sets.
  - This is in line with the ”studybase principle” –- that controls should be selected from the population that gave rise to the cases. (Miettinen 1985)

- In matched cohort studies, there is a choice between sampling with or without replacement.
- Sampling without replacement has the benefit that all matched sets are independent, i.e. the same unexposed person cannot be selected again and only appears once.
- Sampling with replacement is a nice analogy to the risk set sampling, meaning that unexposed persons are selected from everyone unexposed in the population still at risk.
- Heide-Jørgensen et al [9] assessed sampling with and without replacement in a register-based setting.
- They reported that sampling without replacement could give **bias in some situations**, but did not give bias if the unexposed were selected in a particular chronological order.
- They also reported that sampling with replacement did **not** cause **bias**.
- **Hence: Sampling of matched cohort studies should consider sampling with replacement, unless deemed unnecessary due to low outcome risks or large sampling frames where the resampling probability is low.**

Matched cohort: In summary

- The matched cohort study design involves **matching exposed to unexposed** on confounding factors.
- Matching occurs **at start of exposure** (start of follow-up).
- The chosen **underlying timescale** is matched for, as well as **additional matching factors**.

- Although matching in cohort studies **will control for confounding of matching factors at baseline**, the analysis should typically **also adjust for matching factors**.
- This is because there is typically confounding during follow-up, censoring or competing risks which may distort the matching balance (which is at start of follow-up).

- **Stratified Cox regression** with stratification on matched sets is an efficient method to adjust for matching, rather than adjusting for matching factors as covariates.

- In comparison to a full cohort analysis, cohort matching **sometimes improves the statistical efficiency**, but **improved efficiency is not guaranteed**. The main reason for cohort matching is control for confounding, interest in causal effect, cost-efficiency or computational efficiency.

References. Matched cohort

Rothman KJ, Greenland S, Lash TL. Modern Epidemiology, 3rd edition, book chapter 11: Design Strategies to Improve Study Accuracy.

Sjölander A, Greenland S (2013). Ignoring the matching variables in cohort studies – when is it valid and why? Statist. Med. 2013, 32 4696–4708. DOI: 10.1002/sim.5879. https://pubmed.ncbi.nlm.nih.gov/23761197/

Sjölander A, Johansson ALV, et al. (2012). Analysis of 1:1 Matched Cohort Studies and Twin Studies, with Binary Exposures and Binary Outcomes. Statistical Science 2012, Vol. 27, No. 3, 395–411. DOI: 10.1214/12-STS390

Cummings, McKnight B, Greenland S. (2003). Matched cohort methods for injury research. Epidemiol Rev 2003;25:43–50

Greenland S, Morgenstern H. (1990). Matching and efficiency in cohort studies. American Journal of Epidemiology 1990; 131(1):151–159. https://pubmed.ncbi.nlm.nih.gov/2293747/

Mansournia MA, Hernan MA, Greenland S. (2013). Matched designs and causal diagrams. International Journal of Epidemiology 2013;42:860–869. doi:10.1093/ije/dyt083. https://pubmed.ncbi.nlm.nih.gov/23918854/

Heide-Jørgensen U, Adelborg K, Kahlert J, Sørensen HT, Pedersen L. (2018). Sampling strategies for selecting general population comparison cohorts. Clinical Epidemiology 2018 Sep 25;10:1325-1337. doi: 10.2147/CLEP.S164456. eCollection 2018.

Suissa S. (2007). Immortal time bias in observational studies of drug effects. Pharmacoepidemiol Drug Saf. 2007;16(3):241–249.

Brazauskas R, Logan BR. (2015). Observational Studies: Matching or Regression? Biol Blood Marrow Transplant. 2016 Mar;22(3):557-63. doi: 10.1016/j.bbmt.2015.12.005. Epub 2015 Dec 19. https://pubmed.ncbi.nlm.nih.gov/26712591/

Miettinen OS. (1985). The "case-control" study: valid selection of subjects. J Chronic Dis. 1985;38(7):543-48.