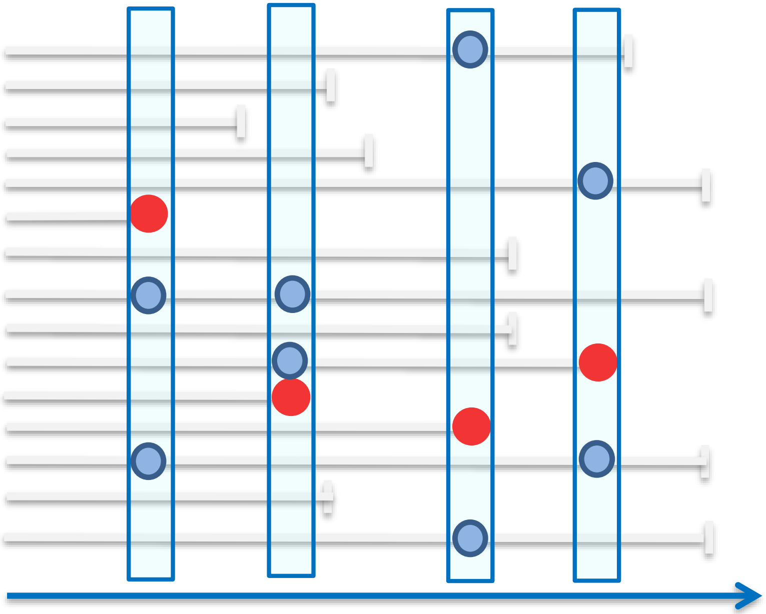
5.2 Breaking the (time) matching

Can we break the matching in nested case-control data?

Several motivations:

reuse of controls?
absolute quantities estimation?

missing data in a set?

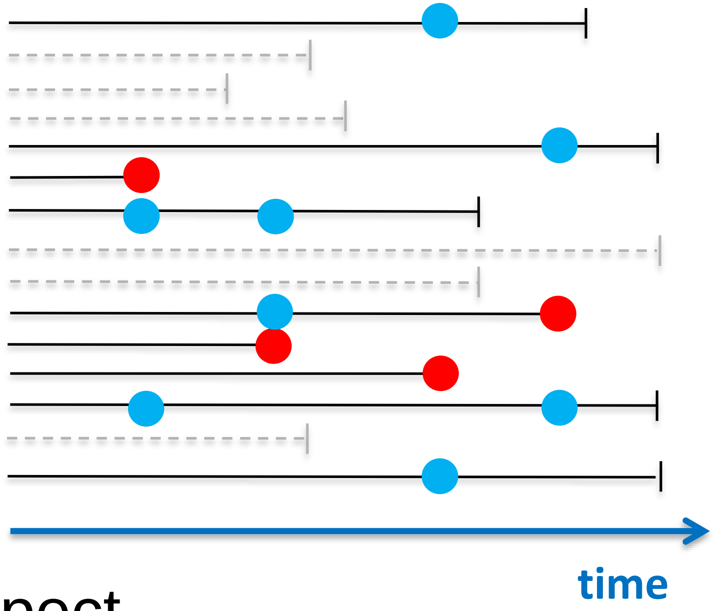


Karolinska Institutet

Objective

- use information from the cases and **all** sampled controls whenever they are at risk

- ”mini-cohort”

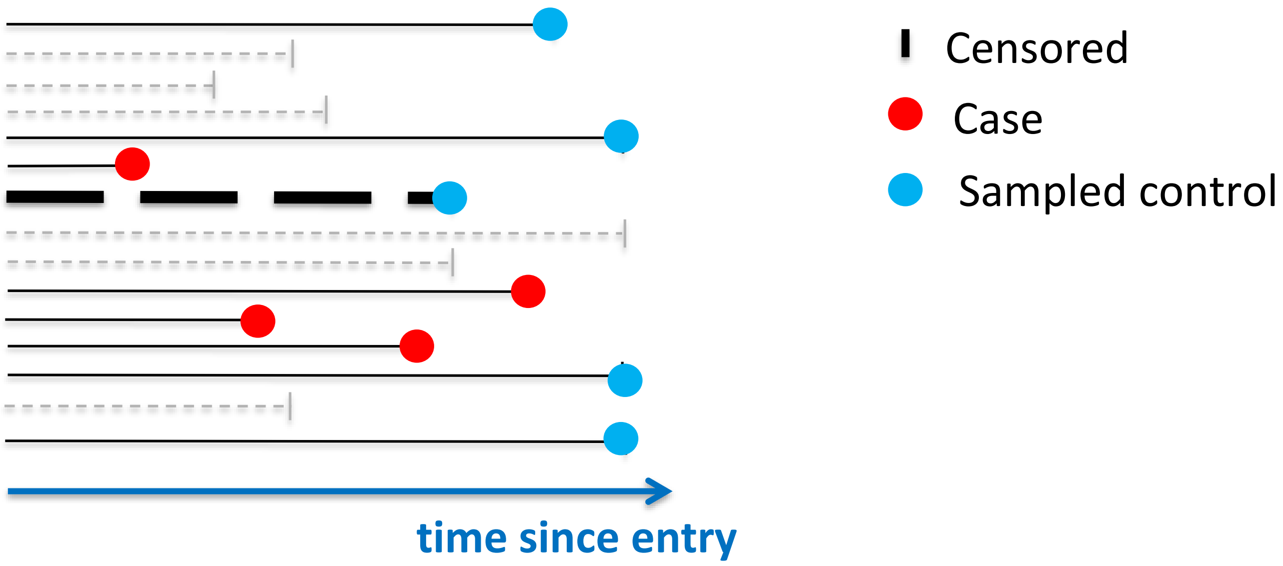


➔ re-introduce the time aspect

➔ If we know the proportion of controls **not** sampled at each event time, we could up-weight those who **were** to be representative of the cohort!

Idea: Upweight the controls by Inverse of the Probability of being Sampled (”IPW” weights)

For each sampled control, we need to find the probability that this particular person was sampled for the study.



IPW weight (inverse probability of being sampled)

- Assume probability =1 for cases (i.e. all cases are in the study)

- For those who do not develop the disease, probability of being sampled as a control depends on:
  - → **The matching variables of the cases**
  - → **No. of potential control ”candidates” at each event time**
  - → **No. of controls to be selected per case**

NOTE: need numbers in cohort at risk at each event time,
and their matching variables

Simpler to calculate probability of ***not*** being selected!

Kaplan-Meier type weight:

Probability individual *i* not sampled at all in the study

$$1 - p_i \quad = \quad \prod_{j, e_i \le t_j \le t_i} \underbrace{\left[1 - \boxed{\frac{m_j}{R_j - 1}} \underline{I(U_j = U_i)}\right]}_{\text{Probability not sampled for case j}}$$

(Samuelsen, *Biometrika* 1997)

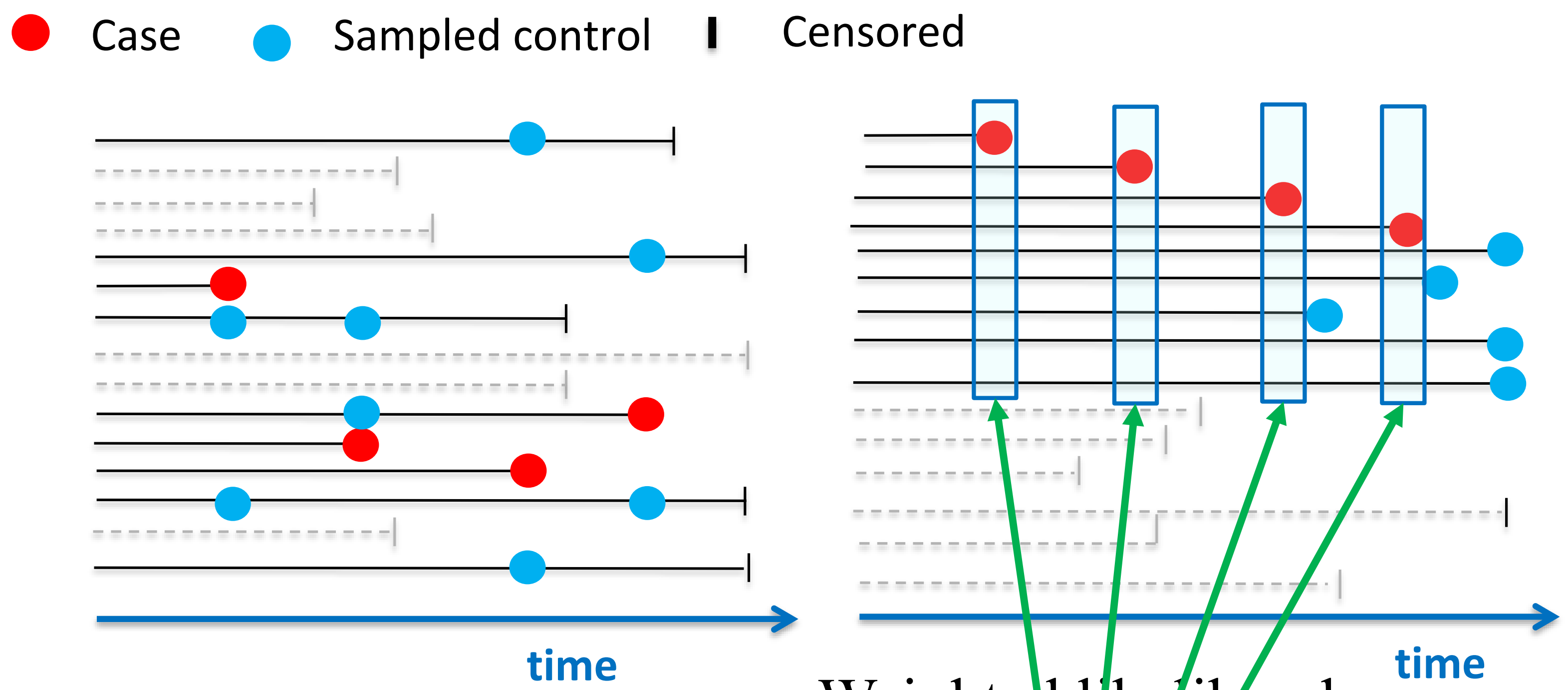
Kaplan-Meier type weight:

$$1 - p_i \quad = \quad \underbrace{\prod_{j, e_i \le t_j \le t_i} \underbrace{\left[1 - \boxed{\frac{m_j}{R_j - 1}}\, \underline{I(U_j = U_i)}\right]}}$$

- e_i and t_i: entry and censoring time of control i
- m_j: number of controls selected for case j
- R_j number of individuals that are still at risk at time t_j
- $I(U_j = U_i)$ matching stratum indicator

w_i= 1 if *i* is a case
w_i= 1/p_i if *i* is a non-case

Weighted Cox regression of the NCC subjects



Weighted likelihood

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

IPW

Solid lines: in NCC sample
Dashed lines: not selected

Compare conditional and weighted analysis

NCC

$$\prod_{t_i} \frac{\exp[\beta X_i + \gamma Z_i]}{\sum_{k \in R^{*}_{i}} \exp[\beta X_k + \gamma Z_k]}$$

Risk set sampled at t_i

Weighted

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

weight

all NCC subjects at risk at t_i

(Borgan and Samuelsen, *Norsk Epi.* 2003)

What about resampled individuals

A weight is calculated for each sampled *individual*

weighted likelihood run on all *unique individuals* (cases or controls)

- A control that later became a case has weight 1
- A control that was sampled twice is only in the data once!

Advantages of breaking the matching

The (weighted) controls can be used as a comparison group for another outcome/disease of interest in the same cohort

The (weighted) Cox regression:

- Provides estimates for all HR (even for the matching factors)
- allows estimation of the absolute risk

So we will learn how to compute the weights!

First some examples of what can be done by breaking the matching (Lecture 5.3)