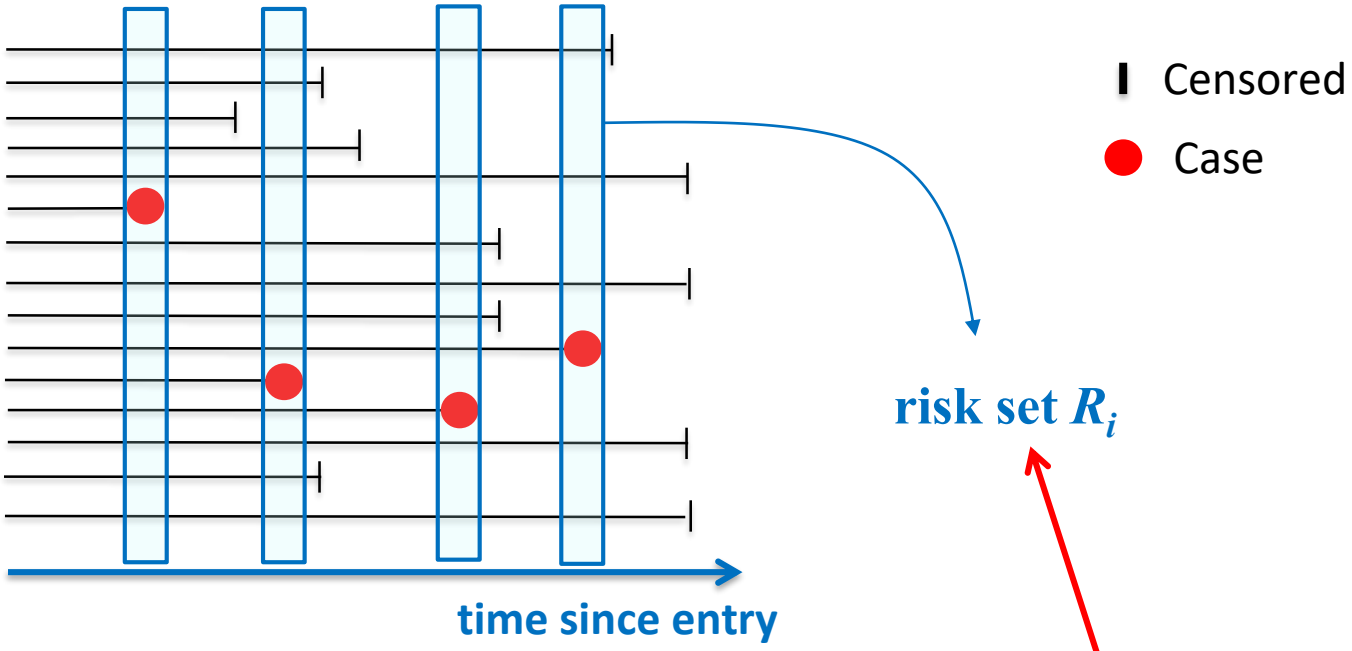
3.3 Matching on time

HR from conditional logistic regression

Cohort design: Cox Regression



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

Do we need the whole cohort?

Do we need the full risk set at each event time?

If the cohort is very large, it seems that we could represent

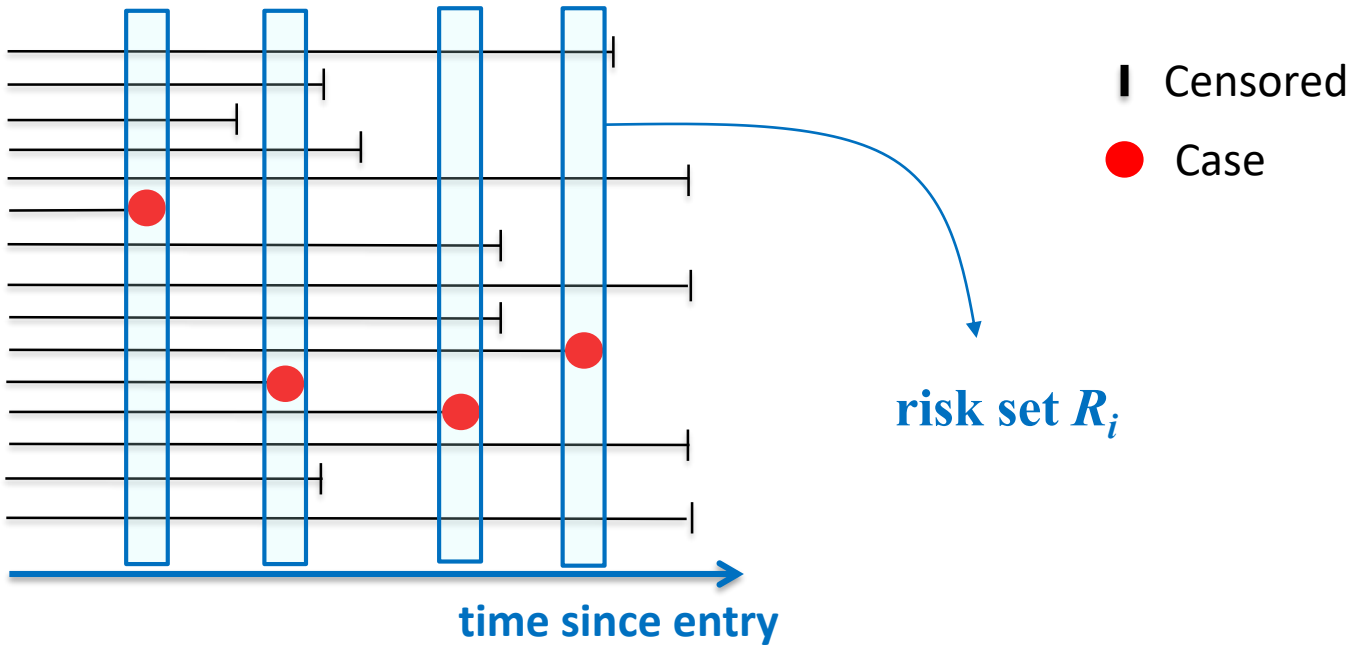
- each risk set by a sub-sample of the “at-risk” individuals at that time point

  (this is the idea of the nested case-control design…. This lecture)

- The experience of the whole cohort from a representative subsample

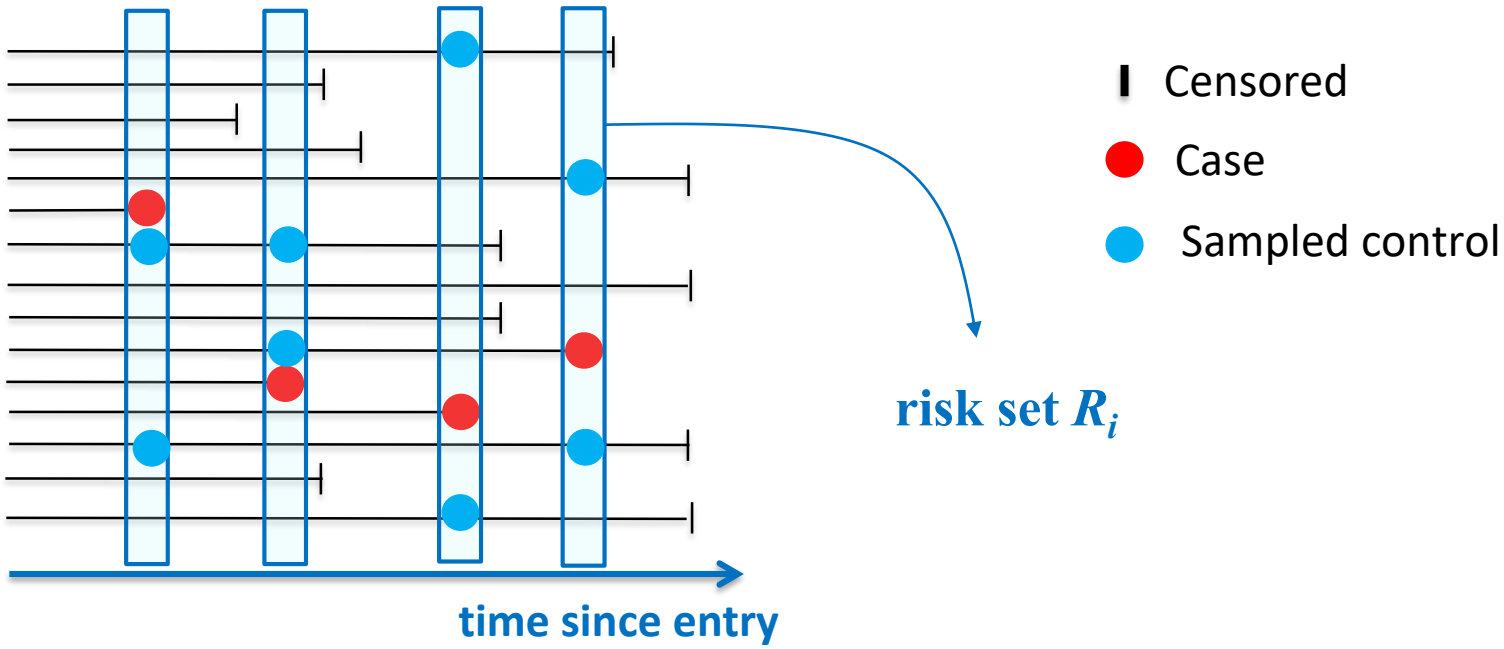
  (this is the idea of the case-cohort design…. later)

Nested case-control design: the concept



- Avoid collecting all cohort data (reduce cost and time)
- *Sample* a few controls from the risk set at each event time, i.e. ”concurrent sampling”

NCC design: the concept (cont.)

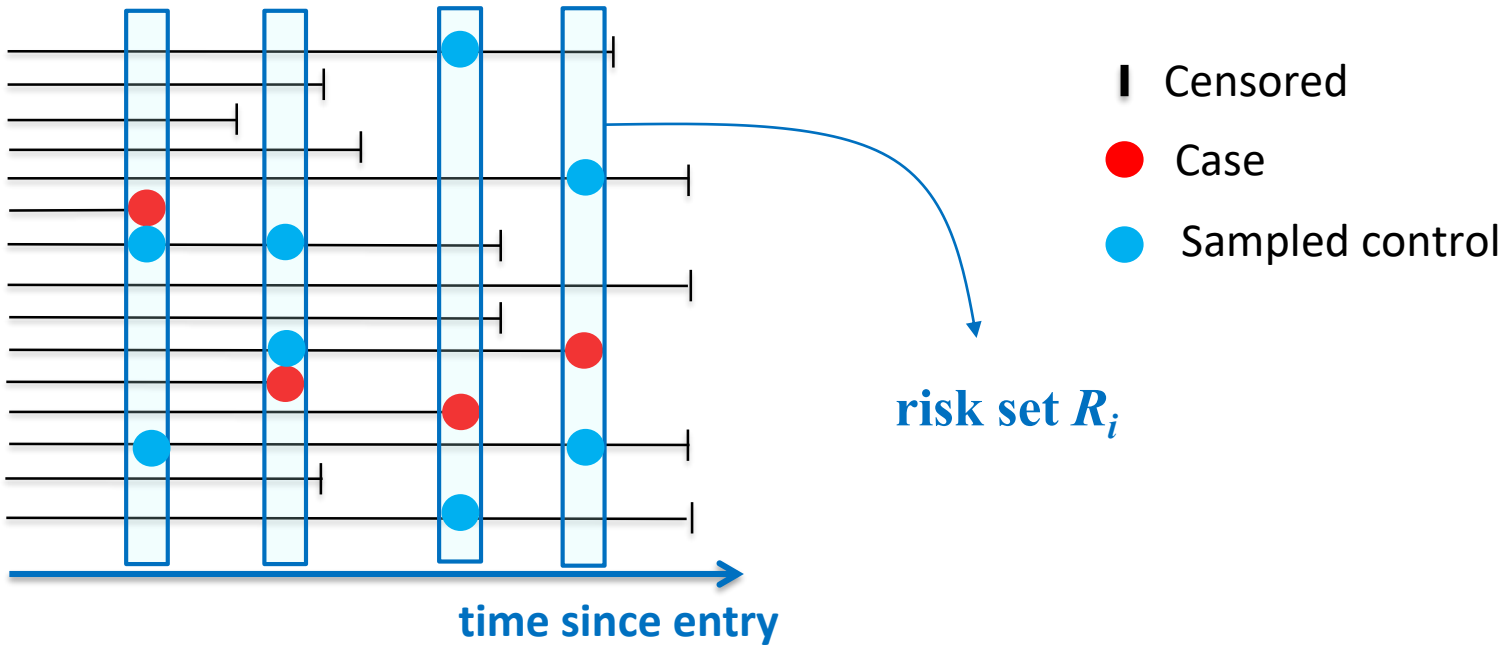


Here, 2 controls are randomly sampled from the risk sets

More controls => better efficiency, usually, 1 to 5 controls

Relative efficiency: **m/(m+1) rule**

NCC design: the concept (cont.)



An individual can be sampled several times

An individual sampled as control can later become a case

NCC design: vocabulary

In this course, a nested case-control study is one that is time-matched as in the previous diagrams, i.e. using

concurrent sampling

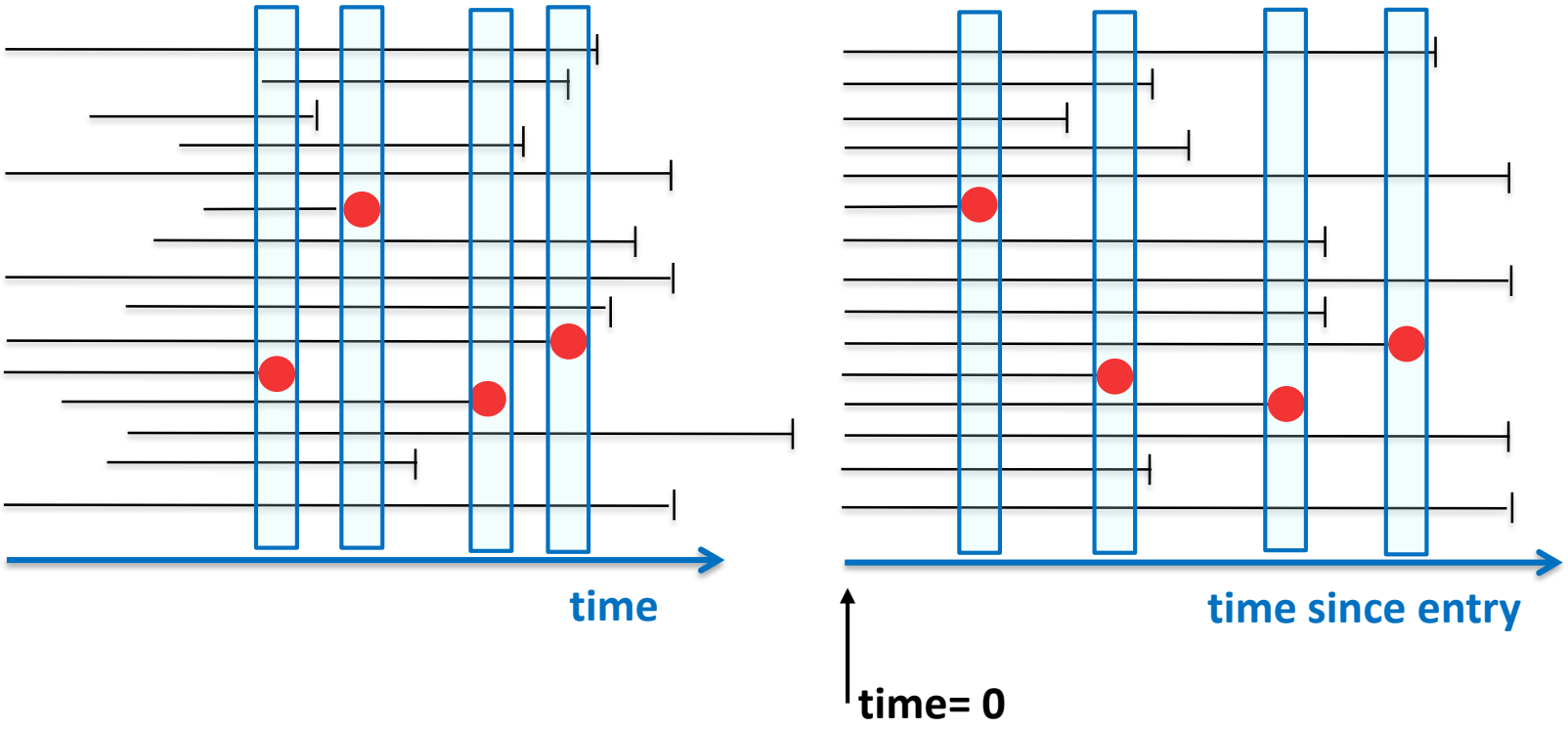
also called incidence density sampling

- In epidemiological papers, if case-control data is sampled within a well defined cohort, the study is described as a *nested case-control* study

All case-control studies can be viewed as *nested* within some population: but we will use the stricter definition of a nested case-control study

NCC design: different risk sets definition

(”study time” may be different than real time)

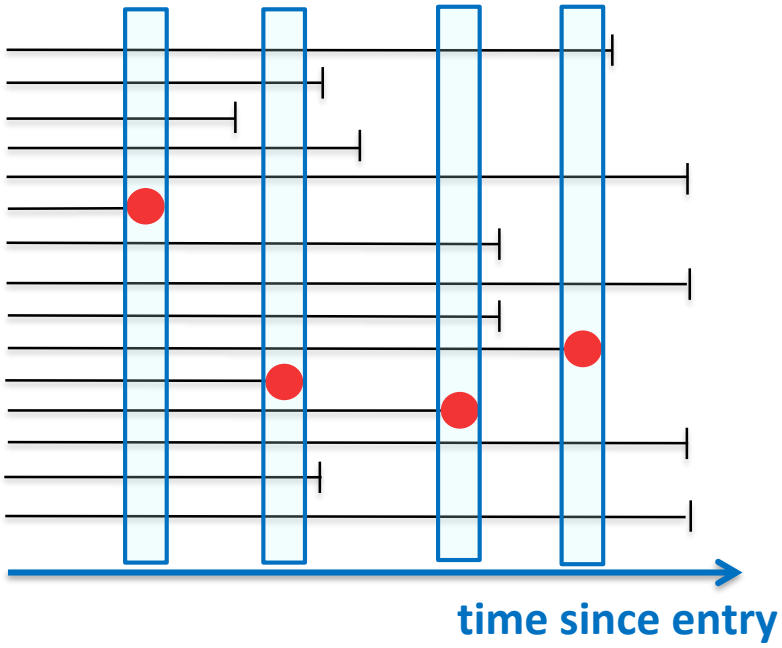


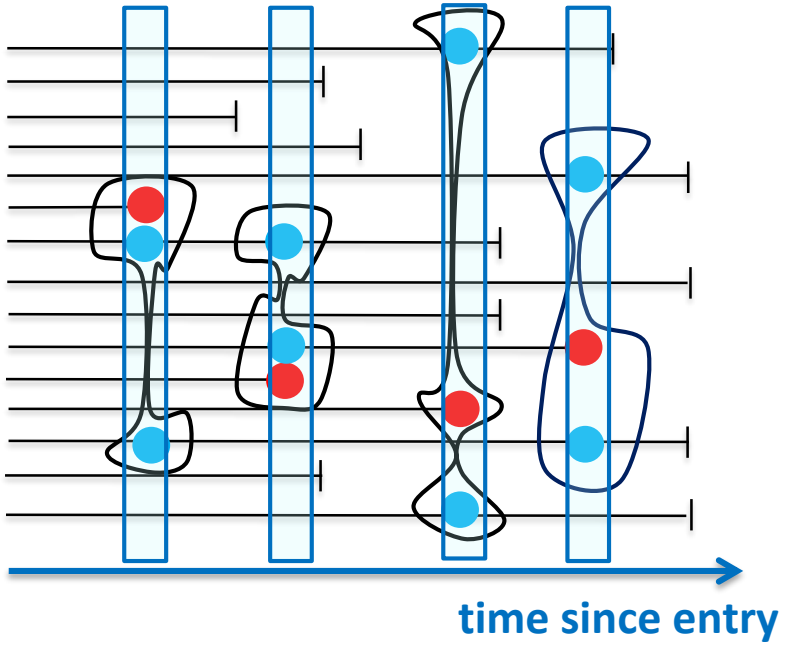
NCC design: analysis

Assume Cox model describing the event in the cohort:

$$h(t|X,Z) = h_0(t)\exp[\beta X + \gamma Z]$$

Cohort vs. NCC:





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

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

Estimating coefficients β and γ

Cohort:

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

risk set

Nested case-control:

$$L(\beta,\gamma) = \prod_{t_i} \frac{\exp[\beta X_i + \gamma Z_i]}{\sum_{k \in R^*_i} \exp[\beta X_k + \gamma Z_k]}$$

sampled risk set

For 1:1 case-control ratio:

Each sampled risk set is just a case-control pair and if we model only exposure (X) effect, the likelihood contribution is:

$$\frac{e^{\beta X1+\gamma Z1}}{e^{\beta X1+\gamma Z1}+e^{\beta X0+\gamma Z0}}$$

where X_1 (Z_1) and X_0 (Z_0) are exposure (confounder) values for case and control,

Note similarity to conditional logistic regression!

Similarity to conditional logistic regression

conditional likelihood is exactly same as for logistic model
(some software packages make it clear that the computations are identical, e.g. PROC PHREG in SAS)

But with time-matching and assuming Cox model,
the β we are estimating is the ln(HR)

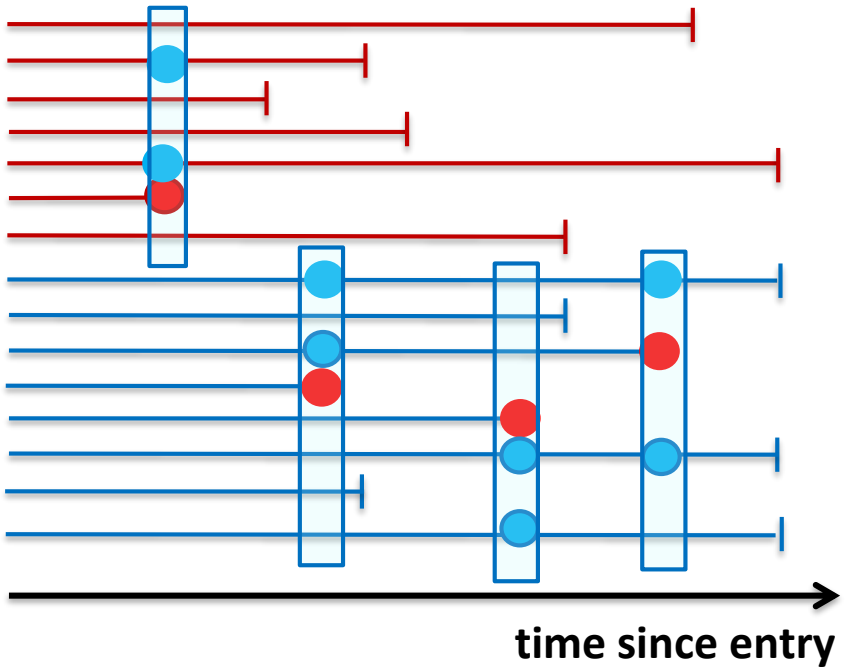
without *time*-matching it is the ln(OR)

More controls => better efficiency, usually, 1 to 5 controls

Relative efficiency: m/(m+1) rule

NCC data with additional matching

One can, in addition to time, match on other variables



Here: the matching variable has 2 strata

NCC data with additional matching (ctd)

- In terms of the sampling:
  The risk sets R_i are reduced in size
- In terms of the analysis:
  Same as before: sampled risk sets R^*_i are used.
  These now include individuals *sharing the same level of the matching variable*
- The matched confounder value(s) are same for all members of a risk set, so cancel from the likelihood.

As before, *assuming a common HR in all matched strata*

Summary: nested case-control design

- Known rule to sample efficient number of controls
- Can estimate HR without following up the cohort
- Simultaneous timing of measurements from cases and controls (may be important for biological specimens)

BUT (standard text-books)

- Only address a question regarding a specific outcome*
- Absolute quantities not easy to estimate*

* possible to solve these issues…