Études cas-témoin, échantillonnage inclusif, études cas-témoins emboîtés

Genève, avril 2012

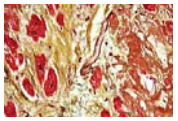
Bernard Cerutti PhD MPH

UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE

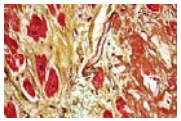
Why a case-control study?

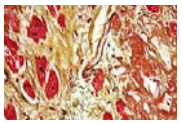
- Rare disease
- Assessment of the exposure is expensive
- Need to inform quickly public health policy makers

The odds ratio



100





75

25

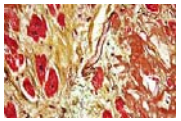
200

60

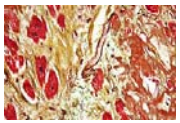
140

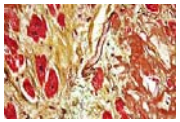
The odds ratio

Odds smokers 75/25 = 3



100





75

25

Odds non smokers 60/140 = 0.43

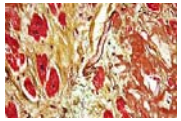
200

60

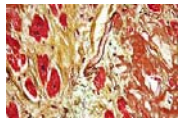
140

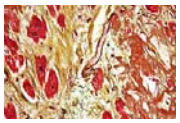
Odds ratio = 3/0.43 = 7

UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE



100





75

25

200

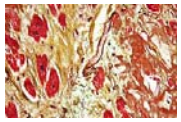
60

140

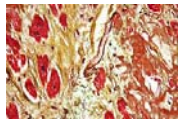
7 smokers have a cardiovascular disease for every smoker without cardiovascular disease

$$\frac{P(C \cap F)}{P(\overline{C} \cap F)} = \frac{P(F/C)\,P(C)}{P(F/\overline{C})\,P(\overline{C})} = \frac{0.75\,P(C)}{0.3\,P(\overline{C})}$$

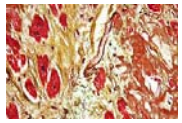
UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE



100



75



25

200

60

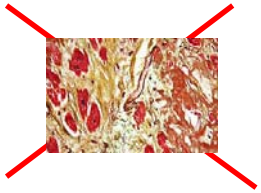
140

If you are smoker the probability that you have a cardiovascular is 7 times higher the the one of a non-smoker

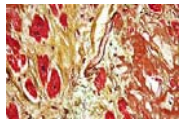
$$\frac{P(C/F)}{P(\overline{C}/F)} = \frac{P(F/C)\,P(C)\,P(F)}{P(F/\overline{C})\,P(\overline{C})\,P(F)} = \frac{0.75\,P(C)}{0.3\,P(\overline{C})}$$

UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE

That is




and



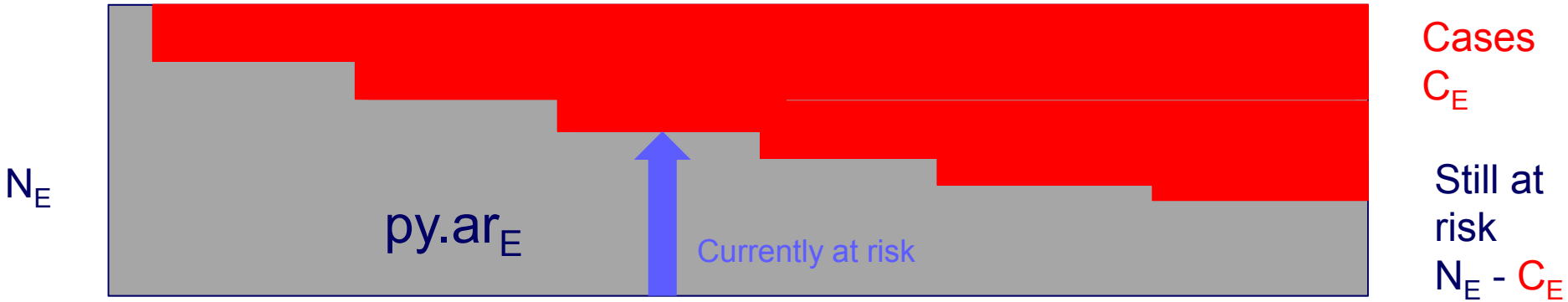

I can bet 7 against 1 that

is a smoker

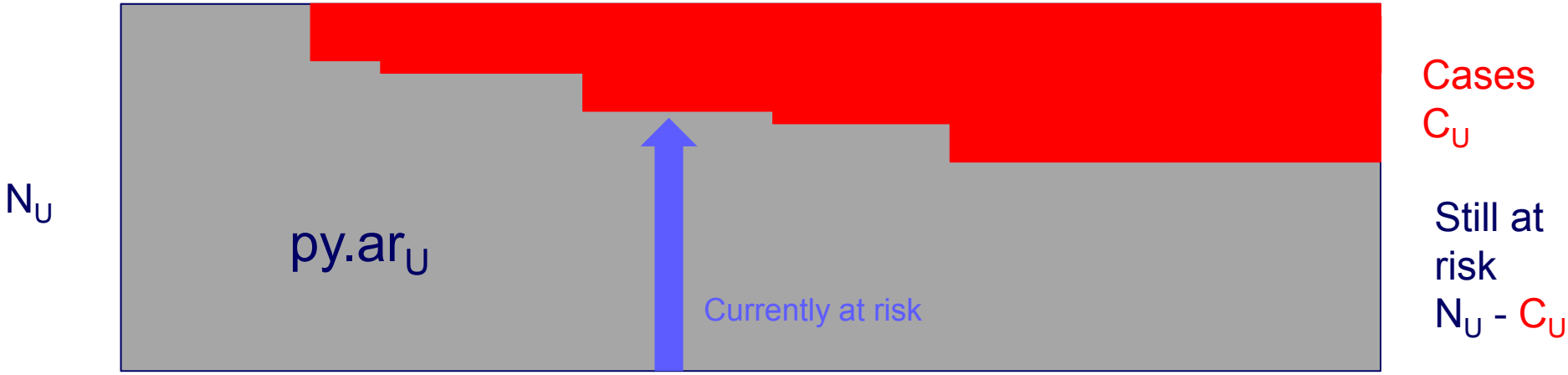
Fixed and dynamic population, stable population

- **Fixed: e.g. birth cohort (Closed)**
- **Dynamic: affected by births, deaths, immigration, … (Open)**
- **Stable population: its composition does not change overtime, neither the exposure**

Exposed population



Unexposed population

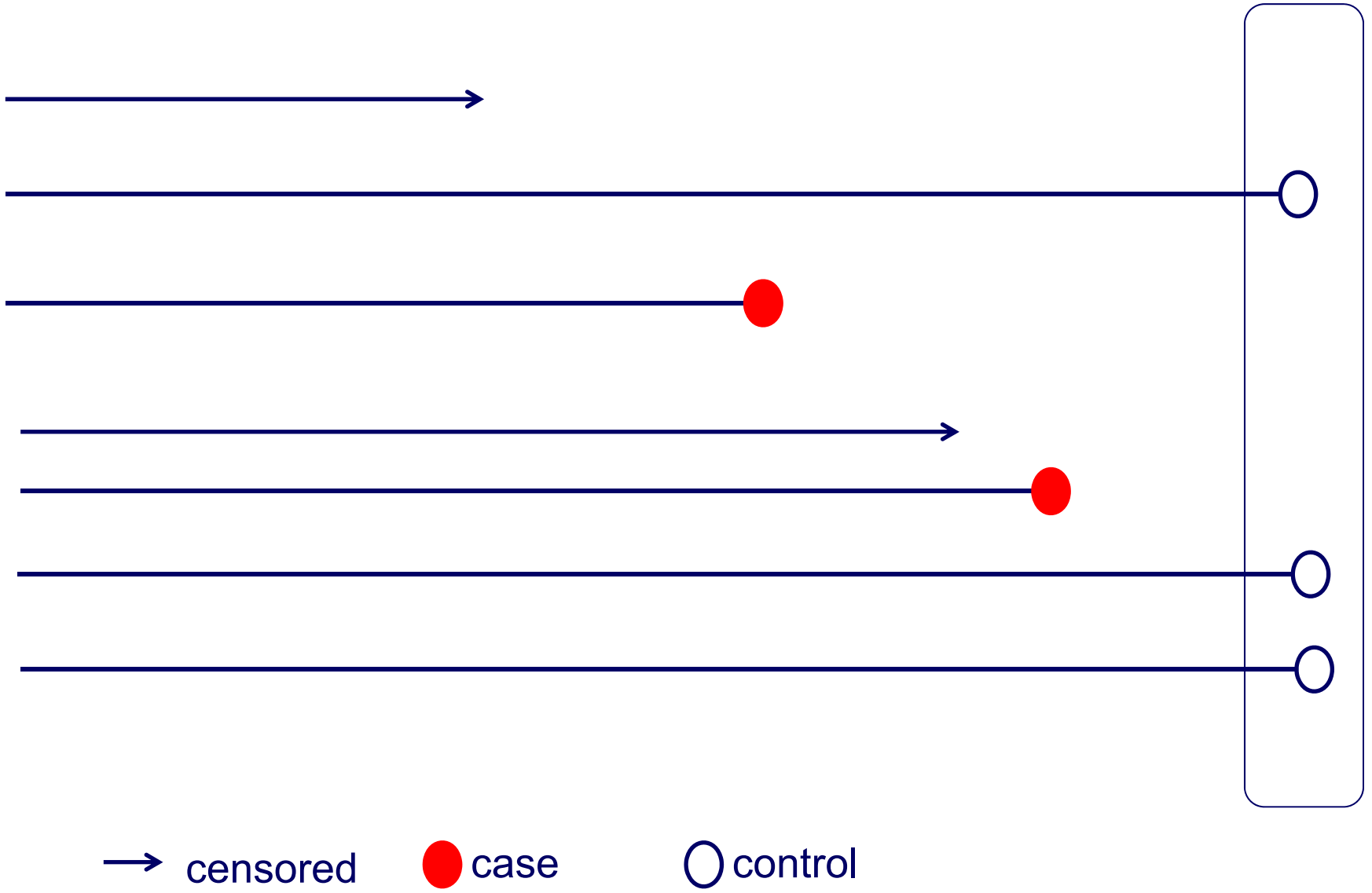


Start of the study

End of the study

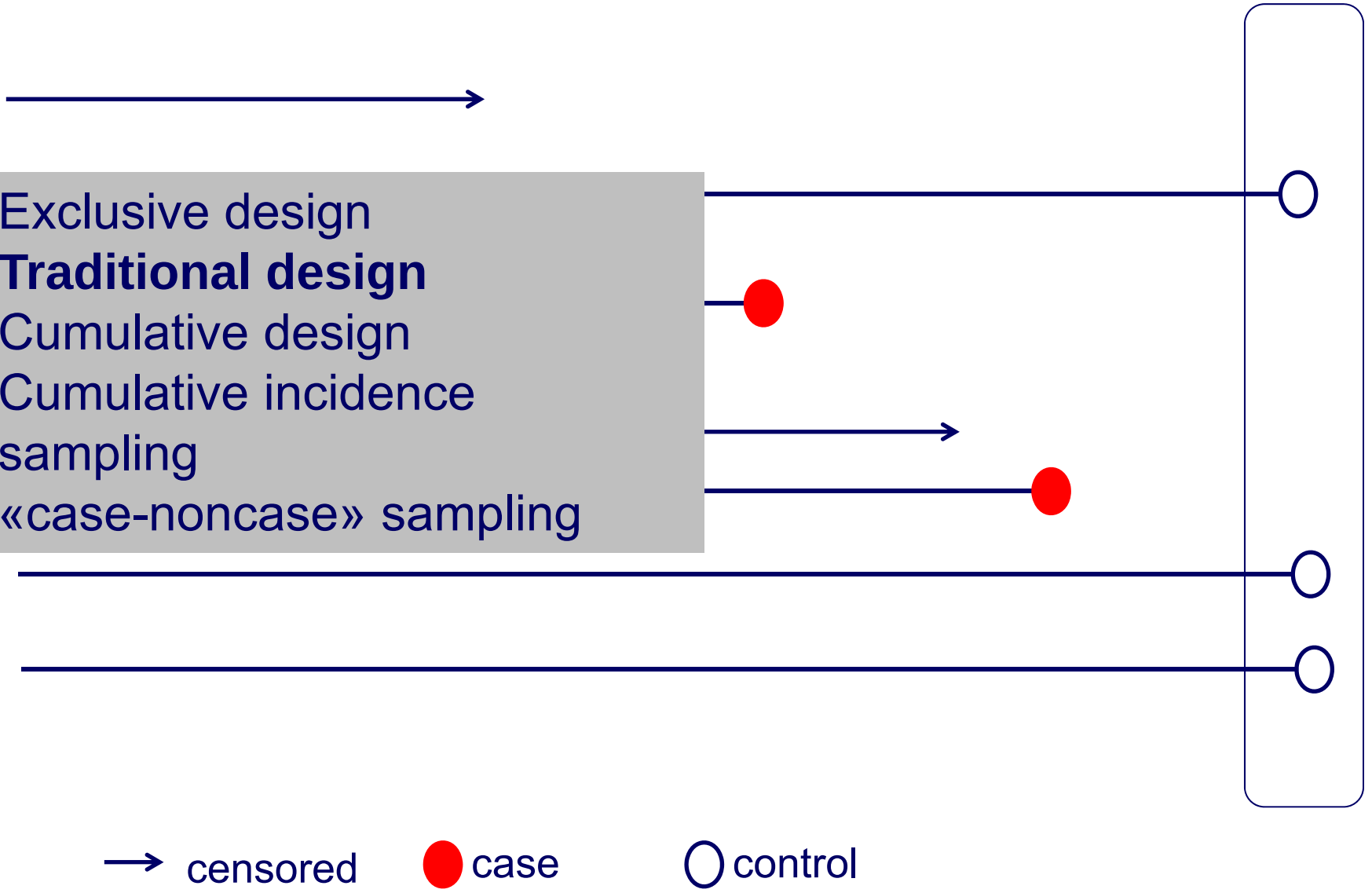
A crucial issue: the approach used to identify the cases and the controls!!

Control selected from the person still free of the disease at the and of the study



Control selected from the person still free of the disease at the and of the study

- Exclusive design
- **Traditional design**
- Cumulative design
- Cumulative incidence sampling
- «case-noncase» sampling



Exposed population

N_E

Cases
C_E

Still at risk
$N_E - C_E$

Unexposed population

N_U

Cases
C_U

Still at risk
$N_U - C_U$

Start of the study

End of the study

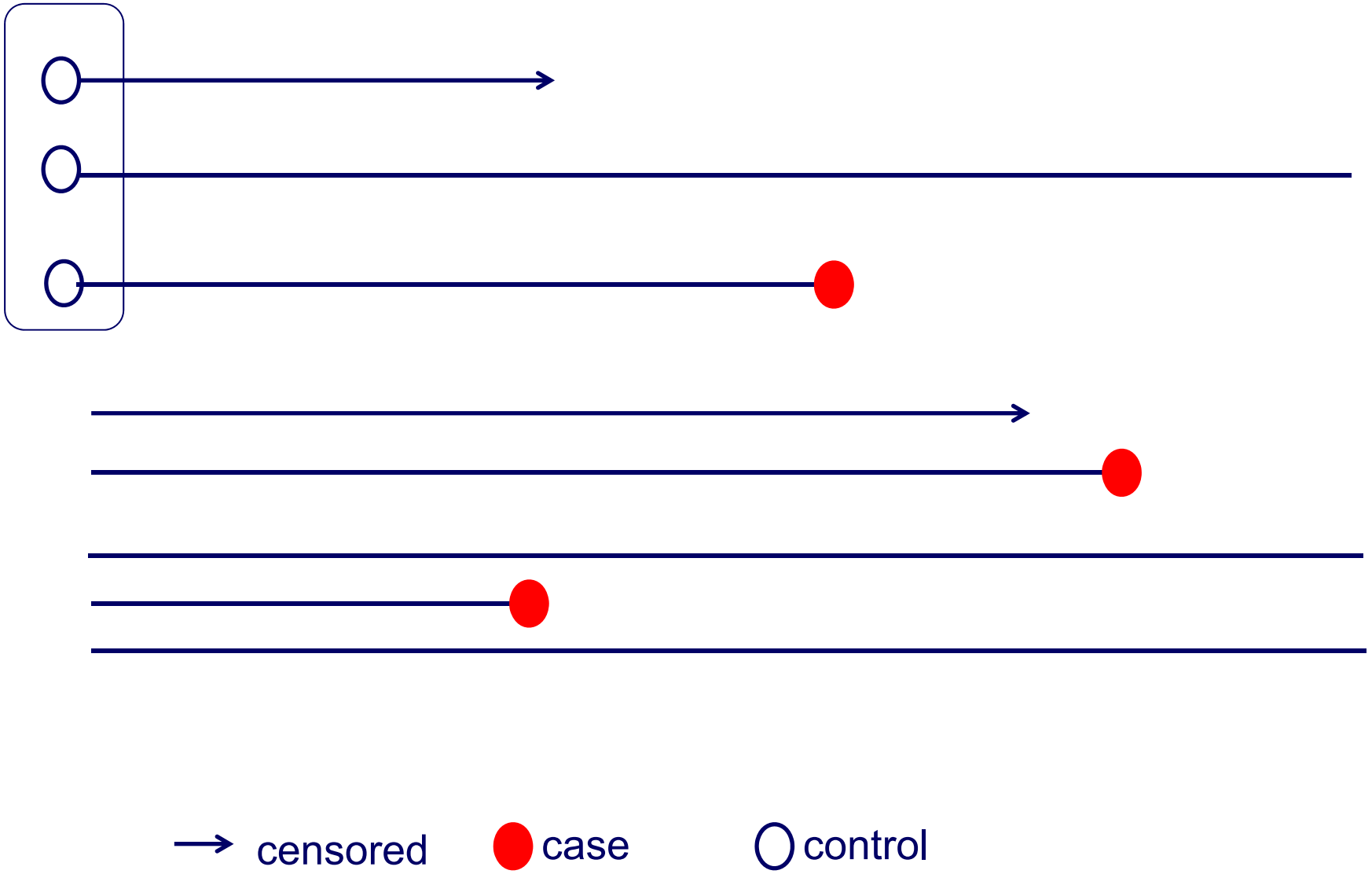
$$\frac{C_E/(N_E - C_E)}{C_U/(N_U - C_U)}$$

Control selected from the person still free of the disease at the end of the study

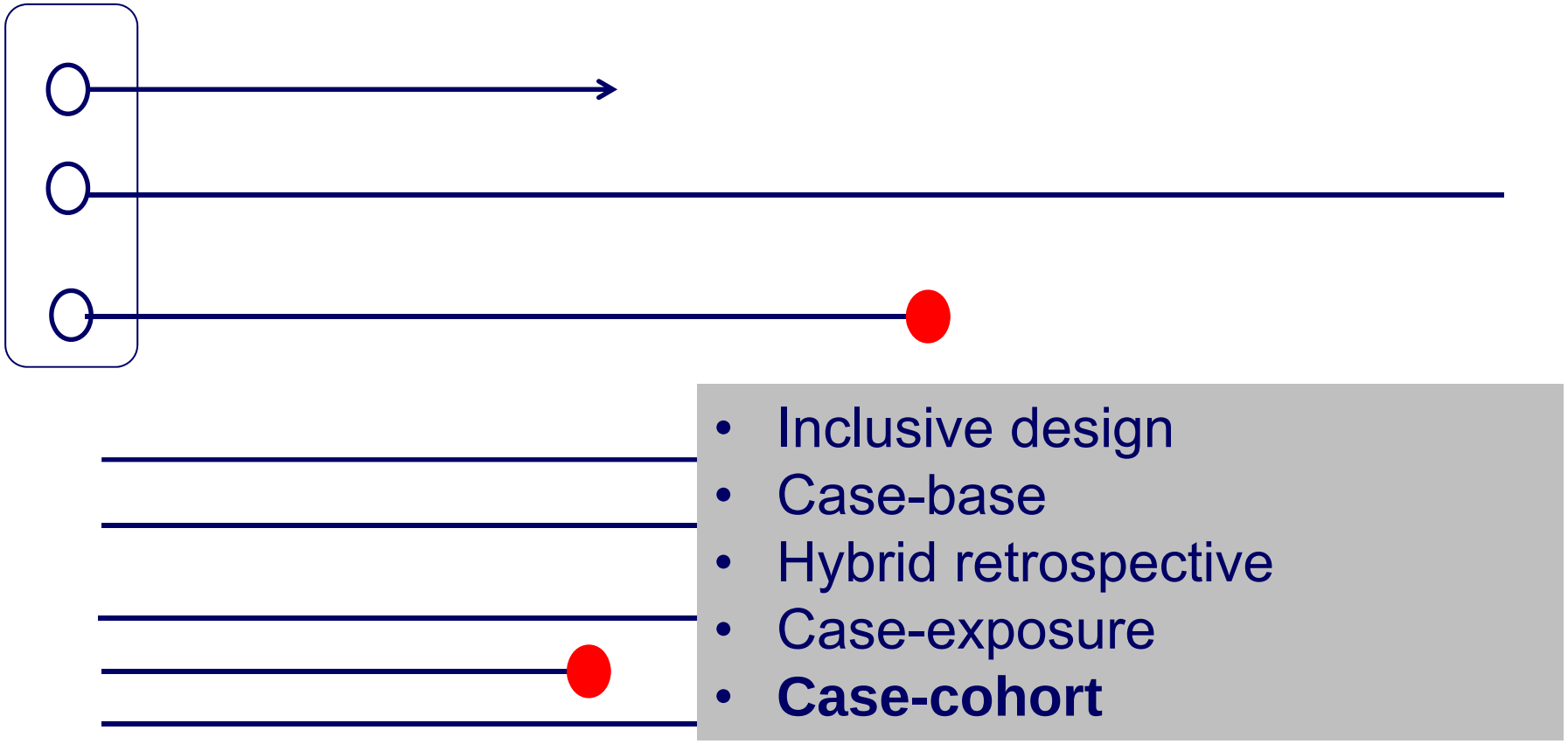
- Many cancer studies
- Congenital studies
- Accidents

- If we have a fixed cohort and the disease is rare (~ incidence below 5%) we can estimate easily the relative risk

Control selected at the beginning of the follow-up period



Control selected at the beginning of the follow-up period



- Inclusive design
- Case-base
- Hybrid retrospective
- Case-exposure
- **Case-cohort**

censored case control

Exposed population

N_E

Cases
C_E

Still at risk
$N_E - C_E$

Unexposed population

N_U

Cases
C_U

Still at risk
$N_U - C_U$

Start of the study

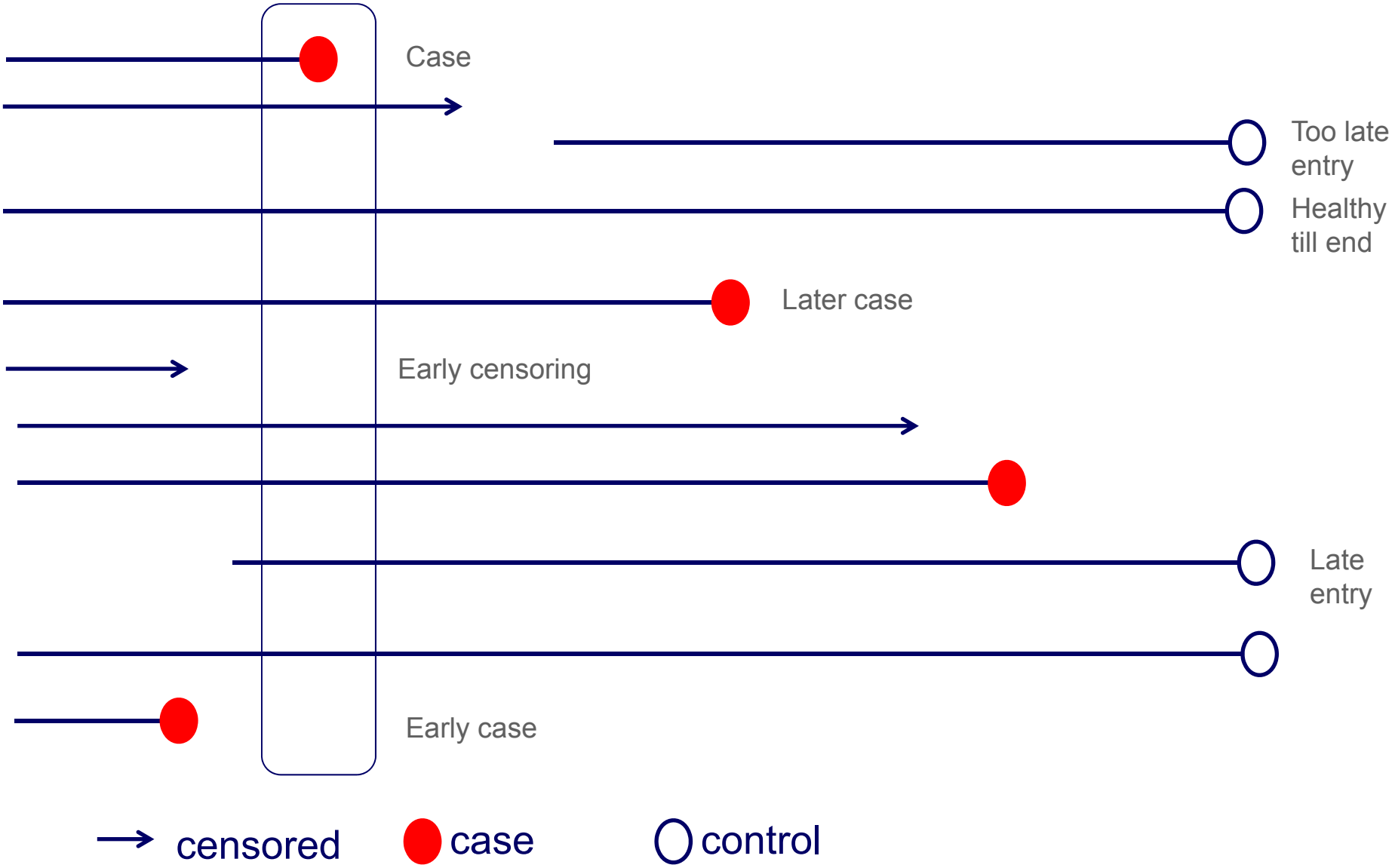
End of the study

$$\frac{C_E / N_E}{C_U / N_U}$$

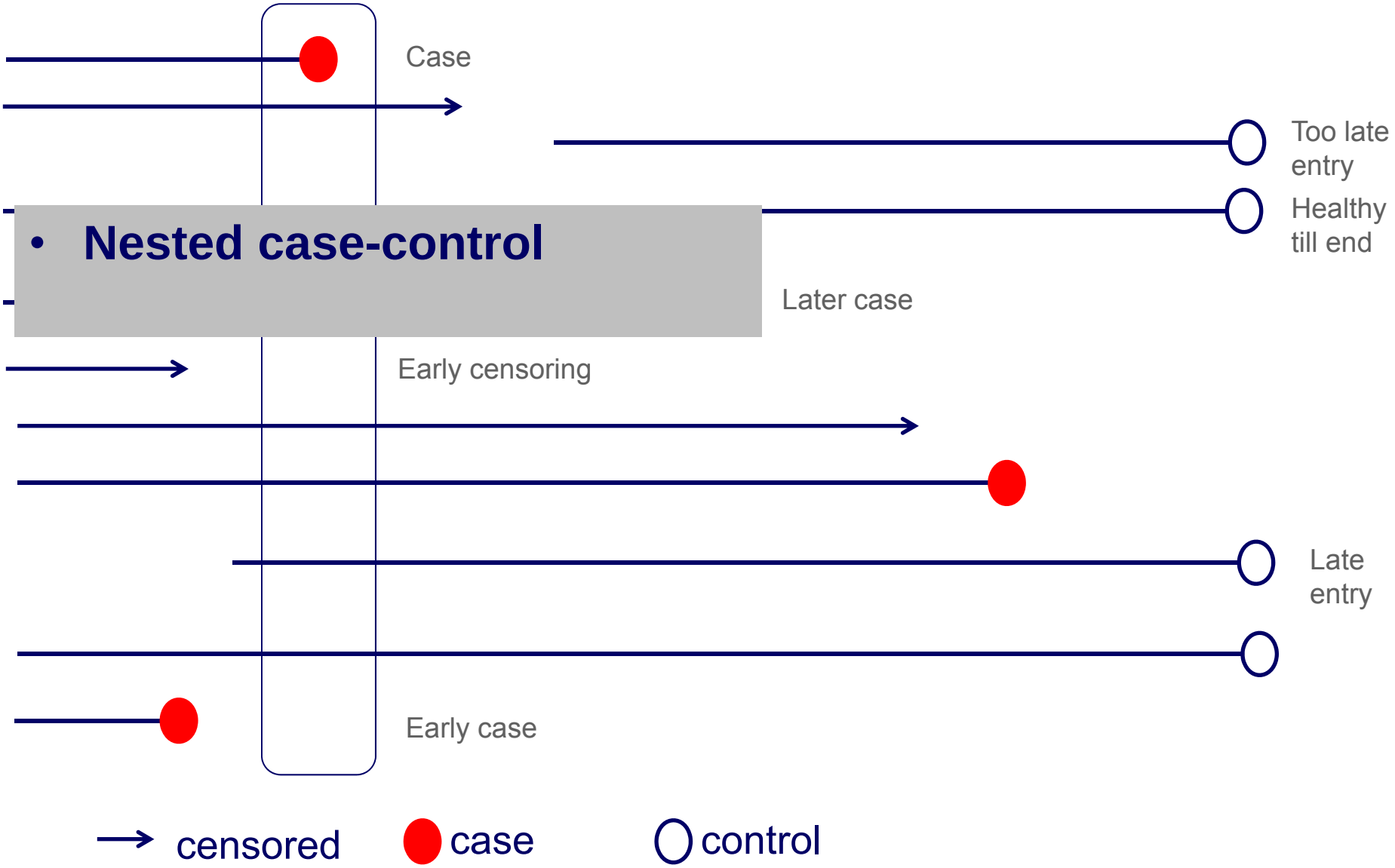
Control selected at the beginning of the follow-up period

- Non-recurrent common disease
- Protective factors which does not affect all exposed equally
- Good especially for multiple outcomes, if measurements of risk factors from stored material remain stable
- Not necessary to obtain the disease history of the selected controls
- The risk ratio is estimable if censoring is unrelated to exposure

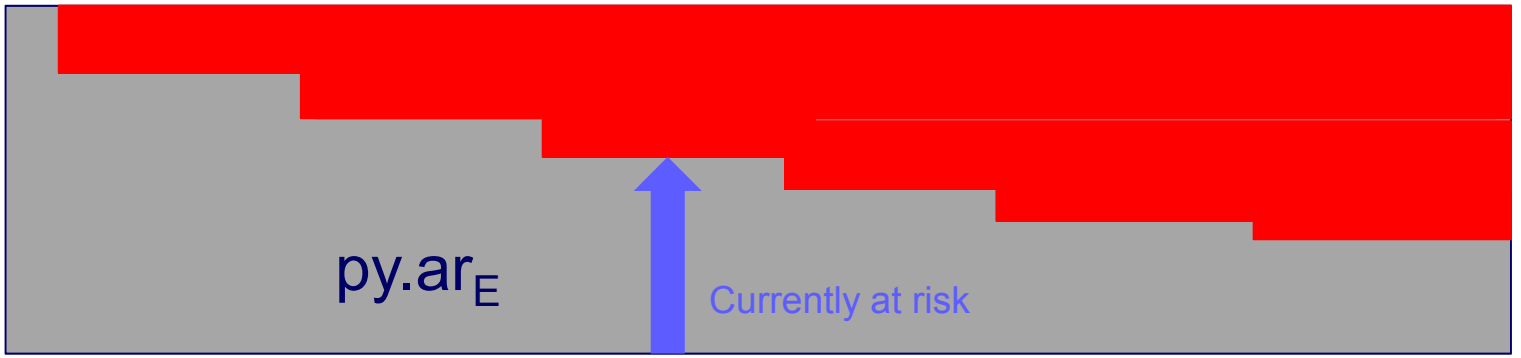
Control drawn during the follow-up



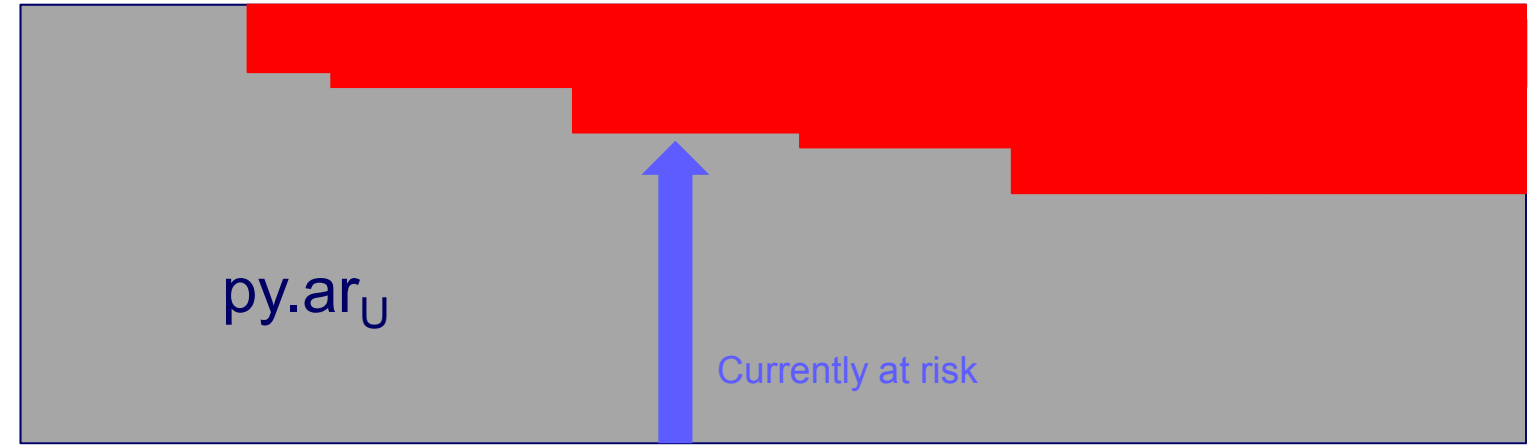
Control drawn during the follow-up



Exposed population



Unexposed population



Start of the study

$$\frac{C_E / C_U}{py.ar_E / py.ar_U}$$

End of the study

Nested-case control

- The **only** logical design in an open population
- Most popular in chronic disease
- Non recurrent common disease with risk/protective factor affecting all exposed equally (eg vaccine with partial protection)
- Recurrent common diseases (diarrhoea, acute respiratory infection)

- About 90% of authors reported having estimated Odds Ratio while they did estimate the Rate Ratio

Matching

- Frequency matching: for cases in a specific stratum, take a set of control from a similar subgroup
- Individual matching: for each case, choose one o more (rarely >5) closely similar controls

- NCC: at least time matching!

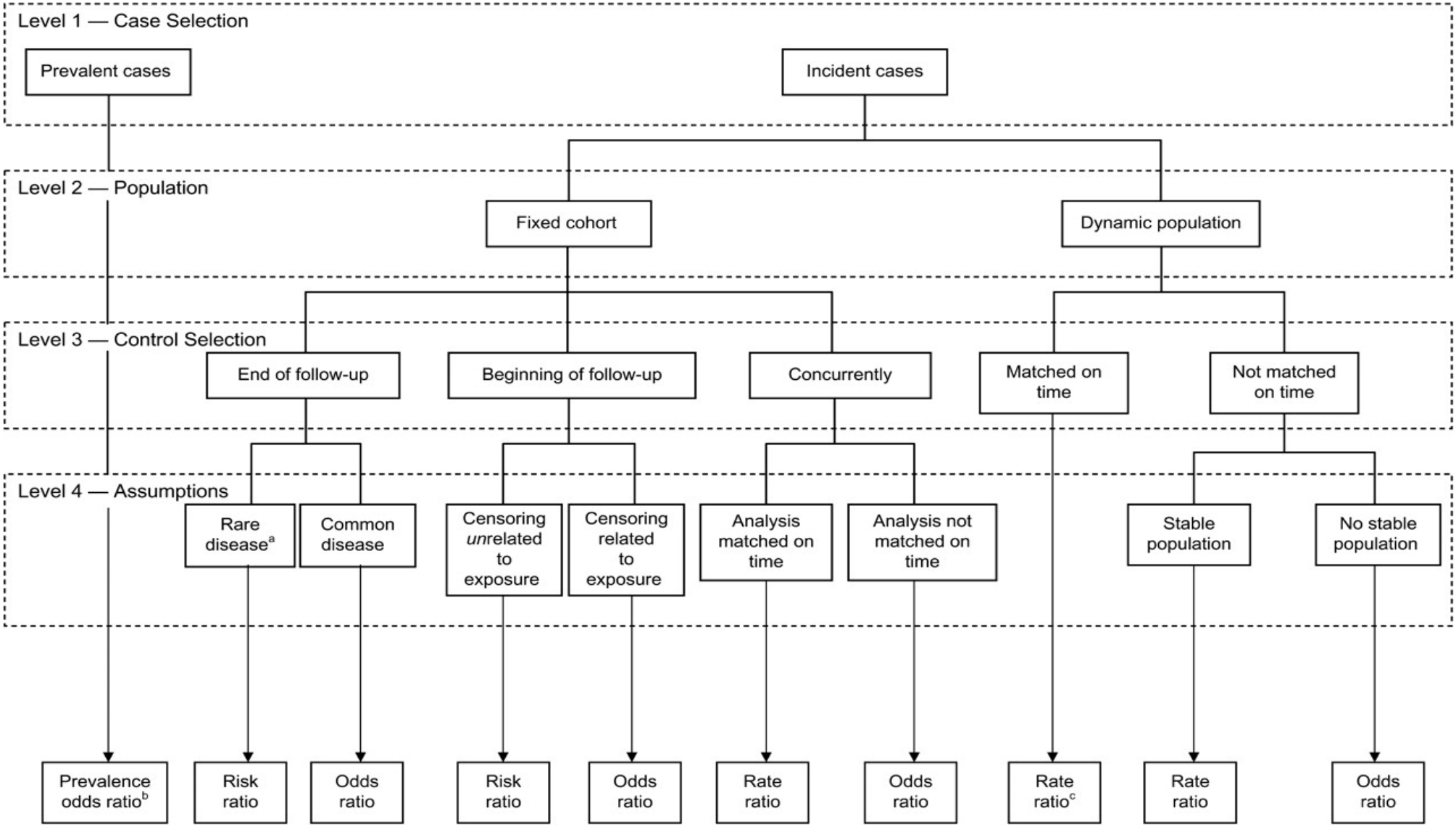
- CC: no matching with cases

Matching

- Increase efficiency if the matching factor are strong risk factors for the disease, and correlated with the main exposure
- Confounding due to poorly quantified factors can be removed by close matching
- Matching on an intermediate variable between exposure and outcome ►bias
- Matching on a surrogate of exposure which is not a true risk factor ►loss of efficiency

The meaning of the odds ratio can depends on the method of selection of the control…

- Are the cases prevalent?
- Are the cases incident?
  - How were the control selected?
    - Population at risk at the beginning
    - Population free of disease at the end
    - Person-time at risk
- Type of the source population
- Sampling strategy
- Underlying assumptions



Knol M J et al. Am. J. Epidemiol. 2008;168:1073-1081

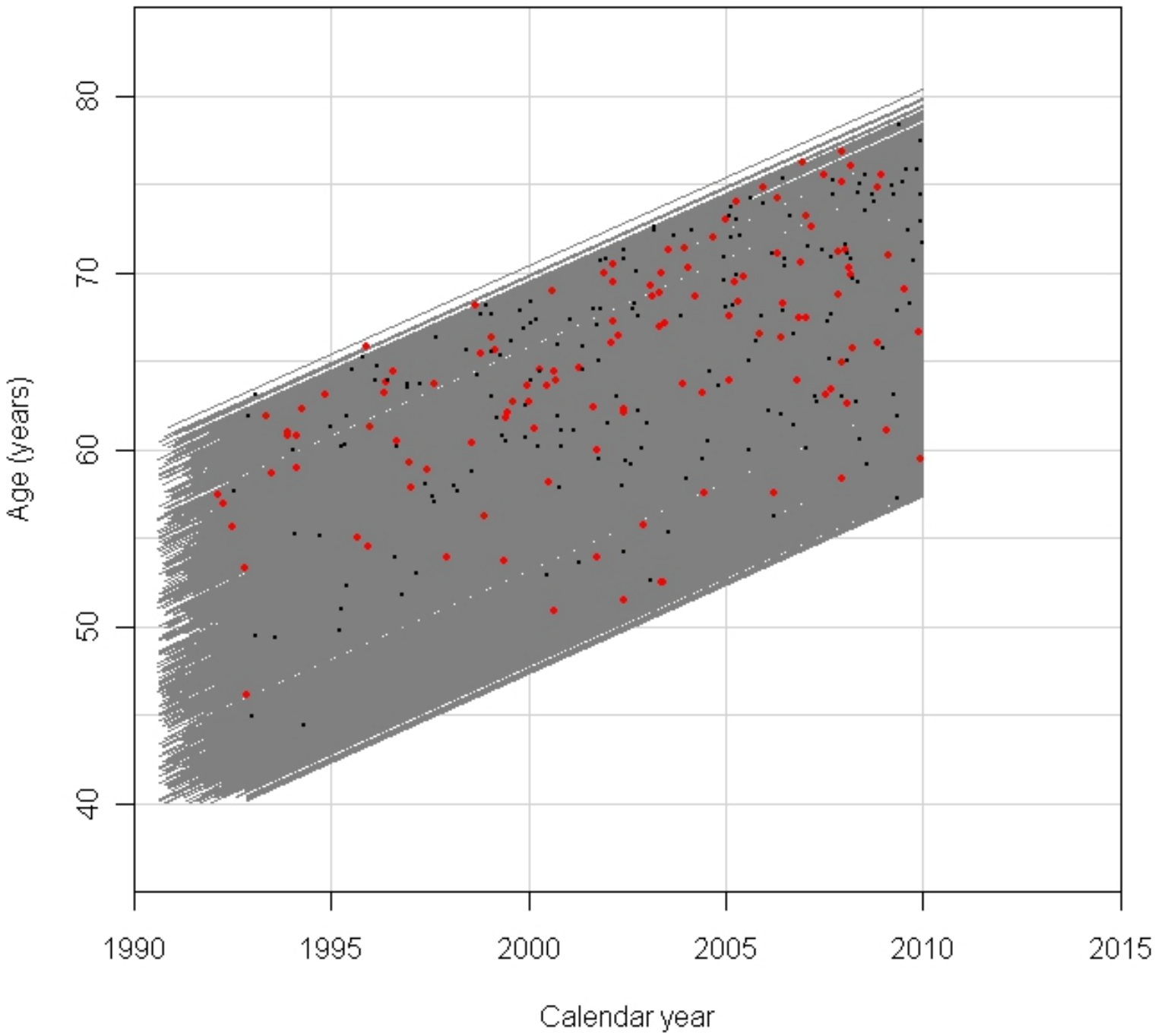
American Journal of EPIDEMIOLOGY

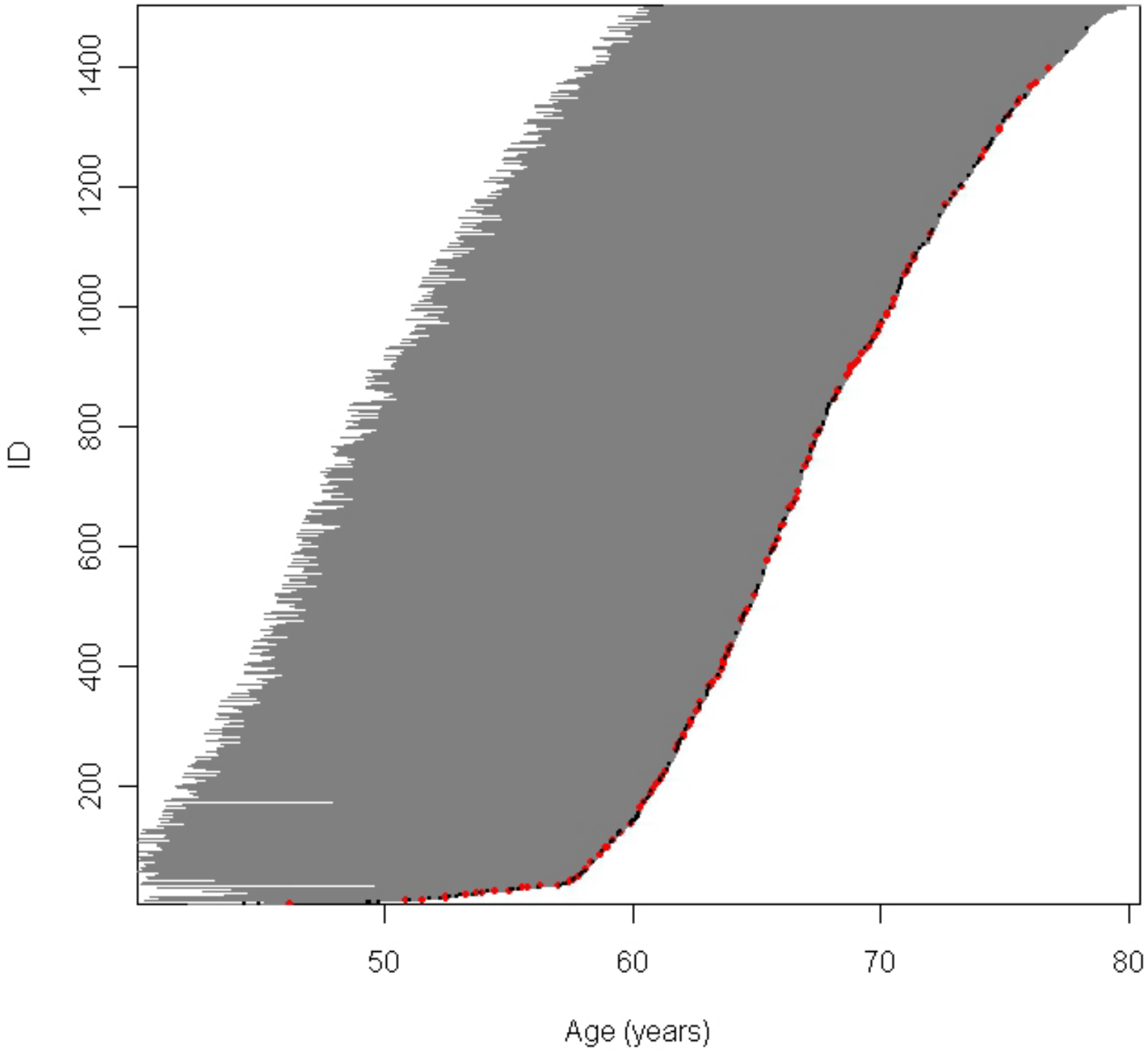
Example

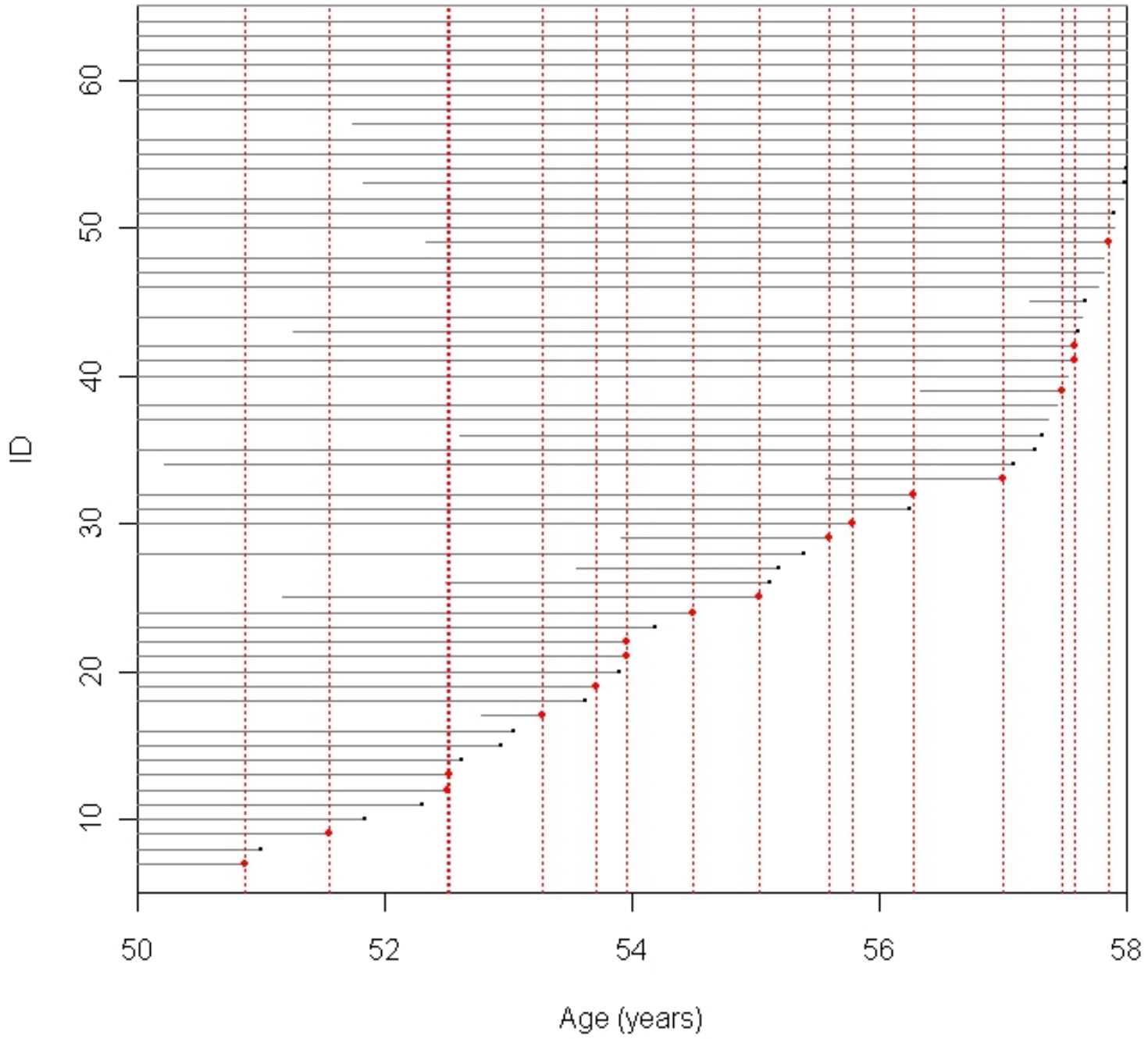
```
> library(Epi)
> library(survival)
> summary(oc)
       id              birth             entry              exit          death          chdeath
 Min.   :   1   1931-02-19:   3   1990-08-18:  12   2009-12-31:1205   Min.   :0.0000   Min.   :0.00000
 1st Qu.: 376   1931-08-24:   3   1991-04-10:  12   2000-01-23:   2   1st Qu.:0.0000   1st Qu.:0.00000
 Median : 751   1933-02-28:   3   1991-04-24:  11   2000-10-04:   2   Median :0.0000   Median :0.00000
 Mean   : 751   1939-04-25:   3   1991-12-18:  11   2001-10-13:   2   Mean   :0.1972   Mean   :0.07995
 3rd Qu.:1126   1941-07-01:   3   1990-11-07:  10   2008-02-09:   2   3rd Qu.:0.0000   3rd Qu.:0.00000
 Max.   :1501   1943-04-16:   3   1991-03-30:  10   2008-03-23:   2   Max.   :1.0000   Max.   :1.00000

> oc$yentry<-cal.yr(oc$entry)
> oc$yexit<-cal.yr(oc$exit)
> oc$ybirth<-cal.yr(oc$birth)
> oc$agentry<-oc$yentry-oc$ybirth
> oc$agexit<-oc$yexit-oc$ybirth

> head(oc)
  id      birth      entry       exit death chdeath   yentry    yexit   ybirth
1  1 1943-02-19 1990-08-14 2009-12-31     0       0 1990.616 2009.997 1943.133
2  2 1934-07-06 1990-08-14 2009-12-31     0       0 1990.616 2009.997 1934.509
3  3 1939-03-05 1990-08-14 2009-12-31     0       0 1990.616 2009.997 1939.172
4  4 1939-07-03 1990-08-14 2009-12-31     0       0 1990.616 2009.997 1939.500
5  5 1935-02-18 1990-08-14 2006-03-13     1       0 1990.616 2006.194 1935.131
6  6 1936-03-07 1990-08-14 2007-06-10     1       0 1990.616 2007.437 1936.179
> oc.lex<-
    Lexis(entry=list(per=yentry,age=agentry),exit=list(per=yexit),exit.status=chdeath,id=id,data=oc)
> summary(oc.lex)
Transitions:
     To
From    0   1  Records:  Events: Risk time:  Persons:
   0 1381 120      1501      120   25280.91      1501
```


Age (years)
40
50
60
70
80
1990
1995
2000
2005
2010
2015
Calendar year


ID
200
400
600
800
1000
1200
1400
50
60
70
80
Age (years)


ID
10
20
30
40
50
60
50
52
54
56
58
Age (years)

Example

```
> oc.lex$agen2<-cut(oc.lex$agentry,br=seq(40,62,1))
> oc.lex$agen2
   [1] (47,48] (56,57] (51,52] (51,52] (55,56]...

> cactrl<-ccwc(entry=agentry,exit=agexit,fail=chdeath,controls=2,match=agen2,
  include=list(id,agentry),data=oc.lex,silent=F)

> head(cactrl)
  Set  Map     Time Fail   agen2   id  agentry
1   1    8 63.93155    1 (47,48]    8 47.72348
2   1 1155 63.93155    0 (47,48] 1155 47.04997
3   1  614 63.93155    0 (47,48]  614 47.35387
4   2   95 66.67762    1 (47,48]   95 47.54278
5   2   11 66.67762    0 (47,48]   11 47.48255
6   2  204 66.67762    0 (47,48]  204 47.56194


> oc.ncc<-merge(cactrl,ocX[,c("id","smok","tchol","sbp")],by.x="Map",by.y="id")
> head(oc.ncc)
  Map Set     Time Fail   agen2 id  agentry smok tchol sbp
1   2  15 64.55305    0 (56,57]  2 56.10678    3  6.55 128
2   8   1 63.93155    1 (47,48]  8 47.72348    2  7.43 154
3  11   2 66.67762    0 (47,48] 11 47.48255    2  5.26 155
4  28  39 66.36824    0 (58,59] 28 58.41752    1  4.56 230
5  33  67 62.76249    0 (53,54] 33 53.01300    4  6.89 127
6  37   8 52.50376    0 (40,41] 37 40.30938    3  5.15 116
```

Example

```
> stat.table(index=list(smok,Fail),contents=list(count(),percent(smok)),margins=T,data=oc.ncc)
 ----------------------------------
          ----------Fail-----------
 smok            0       1   Total
 ----------------------------------
 never          97      31     128
              40.4    25.8    35.6

 ex             55      19      74
              22.9    15.8    20.6

 1-14/d         60      42     102
              25.0    35.0    28.3

 >14/d          28      28      56
              11.7    23.3    15.6


 Total         240     120     360
             100.0   100.0   100.0
 ----------------------------------
```

Example

```
> smok.crncc<-glm(Fail~smok,family=binomial,data=oc.ncc)
> summary(smok.crncc)
Call:
glm(formula = Fail ~ smok, family = binomial, data = oc.ncc)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1774  -0.7704  -0.7447   1.3321   1.6841

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.14072    0.20632  -5.529 3.22e-08 ***
smokex       0.07783    0.33672   0.231 0.817206
smok1-14/d   0.78405    0.28817   2.721 0.006513 **
smok>14/d    1.14072    0.33763   3.379 0.000729 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 458.29  on 359  degrees of freedom
Residual deviance: 441.87  on 356  degrees of freedom
AIC: 449.87

Number of Fisher Scoring iterations: 4

> round(ci.lin(smok.crncc,E=T)[,5:7],3)
            exp(Est.)  2.5% 97.5%
(Intercept)     0.320 0.213 0.479
smokex          1.081 0.559 2.091
smok1-14/d      2.190 1.245 3.853
smok>14/d       3.129 1.614 6.065
```

Example

```
> m.clogit<-clogit(Fail~smok+sbpgrp+cholgrp+strata(Set),data=oc.ncc)
> summary(m.clogit)
Call:
coxph(formula = Surv(rep(1, 360L), Fail) ~ smok + sbpgrp + cholgrp +
    strata(Set), data = oc.ncc, method = "exact")
  n= 360, number of events= 120
                     coef exp(coef)  se(coef)      z Pr(>|z|)
smokex           0.007656  1.007685  0.365587  0.021  0.98329
smok1-14/d       0.673439  1.960970  0.296626  2.270  0.02319 *
smok>14/d        1.139278  3.124510  0.359483  3.169  0.00153 **
sbpgrp[130,150) -0.075530  0.927252  0.326639 -0.231  0.81713
sbpgrp[150,170) -0.066652  0.935521  0.342487 -0.195  0.84570
sbpgrp[170,240]  0.936274  2.550460  0.389203  2.406  0.01615 *
cholgrp[5,6.5]   0.125522  1.133740  0.321175  0.391  0.69593
cholgrp[6.5,13]  0.608167  1.837061  0.353258  1.722  0.08514 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                exp(coef) exp(-coef) lower .95 upper .95
smokex             1.0077     0.9924    0.4922     2.063
smok1-14/d         1.9610     0.5100    1.0964     3.507
smok>14/d          3.1245     0.3201    1.5445     6.321
sbpgrp[130,150)    0.9273     1.0785    0.4888     1.759
sbpgrp[150,170)    0.9355     1.0689    0.4781     1.831
sbpgrp[170,240]    2.5505     0.3921    1.1894     5.469
cholgrp[5,6.5)     1.1337     0.8820    0.6041     2.128
cholgrp[6.5,13]    1.8371     0.5443    0.9192     3.671

Rsquare= 0.075   (max possible= 0.519 )
Likelihood ratio test= 28.09  on 8 df,   p=0.0004582
Wald test            = 24.04  on 8 df,   p=0.002253
Score (logrank) test = 27.08  on 8 df,   p=0.0006854
```

Example

```
> round(ci.lin(m.clogit,E=T)[,5:7],3)
                exp(Est.)  2.5% 97.5%
smokex              1.008 0.492 2.063
smok1-14/d          1.961 1.096 3.507
smok>14/d           3.125 1.545 6.321
sbpgrp[130,150)     0.927 0.489 1.759
sbpgrp[150,170)     0.936 0.478 1.831
sbpgrp[170,240]     2.550 1.189 5.469
cholgrp[5,6.5)      1.134 0.604 2.128
cholgrp[6.5,13]     1.837 0.919 3.671
```

References

- **What do case-control studies estimate? Survey of methods and assumptions in published case-control research**; Knol MJ, Vandenbroucke JP, Scott P, Egger M; Am J Epidemiol; 2008;168:1073-81.
- **Case-control designs in the study of common diseases: updates on the demise of the rare disease assumption and the choice of sampling scheme for controls.** Rodrigues L, Kirkwood BR; Int J Epidemiol; 1990;19:205-13.
- **Nested case-control studies and case-control** studies. Läära E, Plummer M; IARC WHO course, Sep 2011.
- **Representation of follow-up**. Cartensen B; IARC WHO course, Sep 2011.