Karolinska Institutet

6.2 Exposure-dependent sampling

exposure-enriched controls, counter-matching

Benefits of outcome-dependent sampling (case-control and extensions)

Large efficiency gains for rare outcome
(3+ controls per case)

For binary outcome

logistic regression gives valid estimate of population OR

Matched sets/pairs (e.g. discordant twins):
Valid OR from conditional logistic regression

For time-to-event outcome:

matching on time and conditional logistic regression gives valid estimate of HR under proportional hazards

Other outcome-dependent sampling case-cohort design (Lecture 6.1)

Selects a ”subcohort” at baseline (to be used as the comparison group) and (usually) all cases during follow-up.

Efficiency similar to nested case-control (similar sample size)

Analysis:

Weighted Cox regression

Weights = 1 for cases

= inverse of sampling fraction for non-cases

valid estimates of population HR and absolute risk

Familiar exposure-dependent sampling

Some simple designs, for example:

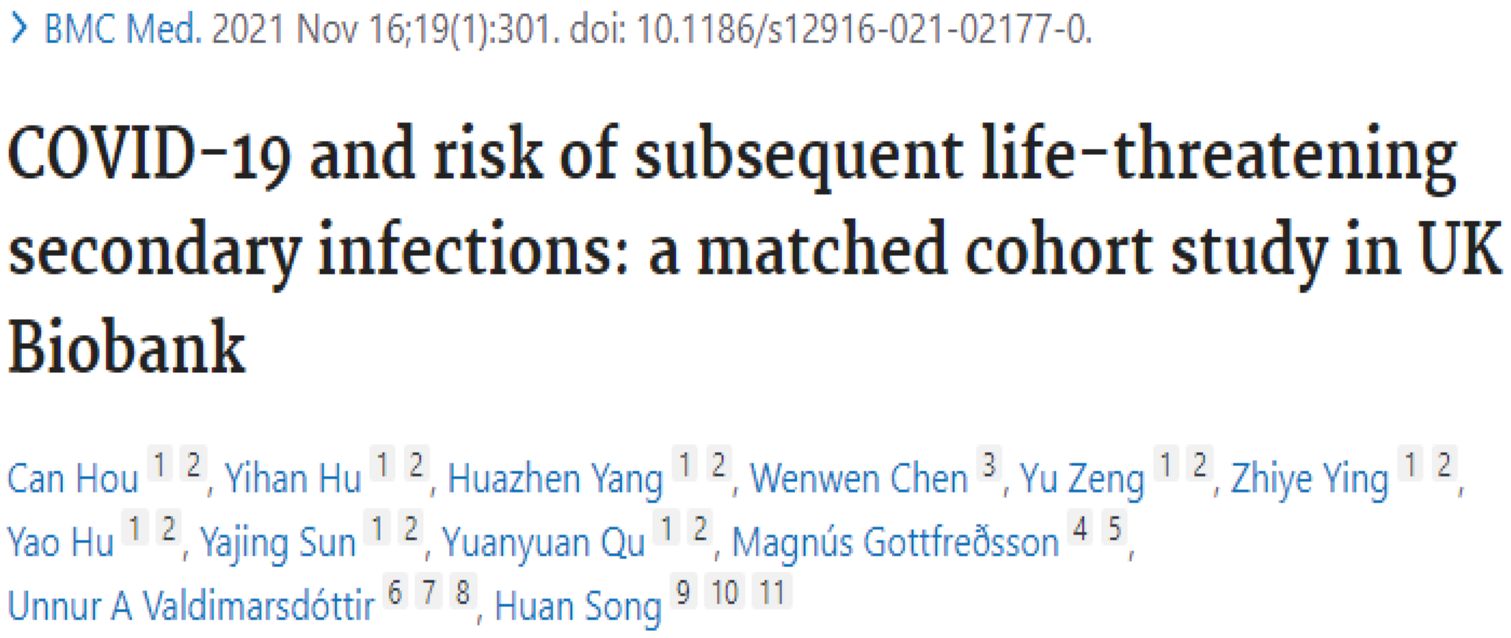
Selection of the numbers of exposed and unexposed individuals (esp. where exposure is rare) in cross-sectional study or at baseline of cohort study

Matched cohort design:

Frequency matching within confounder strata

(matched pairs, 1:1 exposed:unexposed)

Especially useful for studies of *rare exposure*



› BMC Med. 2021 Nov 16;19(1):301. doi: 10.1186/s12916-021-02177-0.

COVID-19 and risk of subsequent life-threatening secondary infections: a matched cohort study in UK Biobank

Can Hou[1,2], Yihan Hu[1,2], Huazhen Yang[1,2], Wenwen Chen[3], Yu Zeng[1,2], Zhiye Ying[1,2], Yao Hu[1,2], Yajing Sun[1,2], Yuanyuan Qu[1,2], Magnús Gottfreðsson[4,5], Unnur A Valdimarsdóttir[6,7,8], Huan Song[9,10,11]

From 445,845 UK Biobank participants, 5151 individuals with a positive test result or hospitalized with a diagnosis of COVID-19 were included in the exposed group.

For each exposed individual, up to 10 unexposed randomly selected matched individuals (n = 51,402).

Cox regression analysis

Less familiar exposure-dependent sampling

…. where exposure-dependent sampling strategies incorporated into familiar design:

Two-stage sampling on a surrogate of exposure √

Exposure-enriched case-control

Exposure density sampling

Counter-matched nested case-control

Exposure-enriched case-control design

Proposed for study of gene-environment interaction

(high efficiency for skewed environmental exposure and rare gene)

Idea: over- (or under-) sample subjects with high/low exposure.

Does *not* need the prevalences required by two-stage design

Straightforward analysis, logistic regression

Motivating Example

Huque et al. *Genetic Epidemiology,* 2016, 40(7), 570-578

Earlier case-control study in 23 villages in Bangladesh to investigate:

- dose-response of water arsenic levels with skin lesions
- interaction with genetic polymorphisms

Investigators had oversampled controls with low exposure (<50µg/l) to “overcome” skewed distribution of arsenic levels

Recall from earlier lectures:

For cohort or cross-sectional data, logistic model is a ”regression model” in the sense that X’s can be fixed/chosen but Y random:

We model P$[Y = 1]$= $\frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}}$

Can compute prevalance from α

But for case-control data, we are modelling P$[\, Y = 1 \;\mid X \,]$ ***conditional on being sampled***

$$= \frac{e^{\alpha*+\beta X}}{1+e^{\alpha*+\beta X}} \qquad \text{where } \alpha^* = \alpha + \log_e \frac{\pi_1}{\pi_0}$$

Returning to the arsenic example

Y= case/control status, skin lesions
X= E (arsenic exposure: well-water and toenail levels)
G (genetic polymorphism in X-ray repair gene (XRCC1 Arg194Trp)
GE (interaction)

Assume population model:

$$\text{Logit } [Y = 1|E, G] = \alpha + \beta_E E + \beta_G G + \beta_{GE} GE$$

For the sampled data, the model is

$$\text{Logit } [Y = 1|E, G, S = 1] = \alpha^* + \beta_E E + \beta_G G + \beta_{GE} GE$$

where S=1 depends on case/control status *and high/low exposure*

Using Bayes Theorem as before..

Logit $[Y = 1|E, G, S = 1]$ = $\alpha + \log_e \frac{\pi_{1H}}{\pi_{0H}} + \beta_E E + \beta_G G + \beta_{GE} GE$ if X>=50

Logit $[Y = 1|E, G, S = 1]$ = $\alpha + \log_e \frac{\pi_{1L}}{\pi_{0L}} + \beta_E E + \beta_G G + \beta_{GE} GE$ if X<50

Where the π_1 and π_0 terms are the case and control sampling probabilities, in the high π_{1H} and π_{0H} and low (π_{1L} and π_{0L})exposure groups.

So, using low exposure as reference, the model can be written:

Logit $[Y = 1|E, G, S = 1]$ = $\alpha^* + \beta_{HL}{}^* \mathrm{I}\ (\mathrm{E} > 50) + \beta_E E + \beta_G G + \beta_{GE} GE$

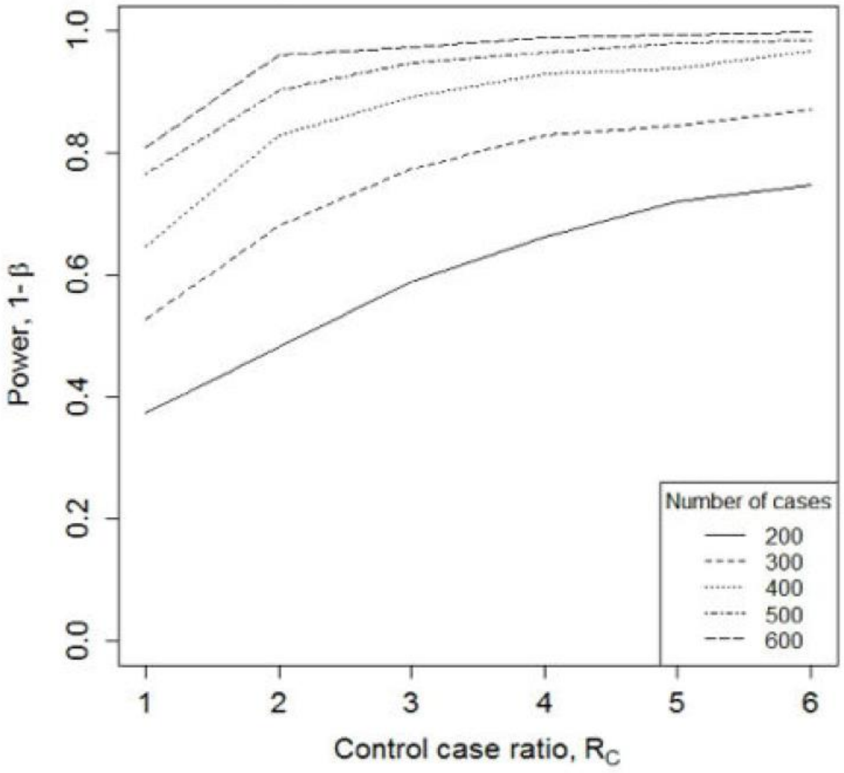
where $\alpha^* = \alpha + \log_e \frac{\pi_{1L}}{\pi_{0L}}$ and β_{HL} is the difference $\log_e \frac{\pi_{1H}}{\pi_{0H}} - \log_e \frac{\pi_{1L}}{\pi_{0L}}$

→ straightforward logistic regression!

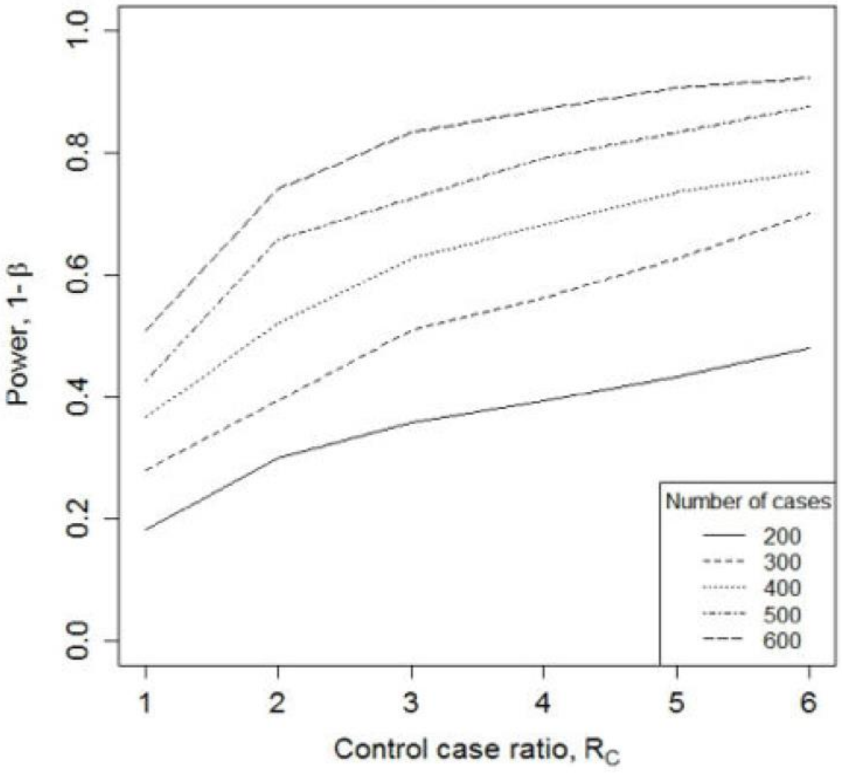
Power of EECC depends on:

- Exposure distribution (asymmetry)
- Ratio of high to low exposed persons in the sample
- Case:control ratio
- gene frequency

3(a): Power associated with Control-case ratio for a given number of cases using EECC



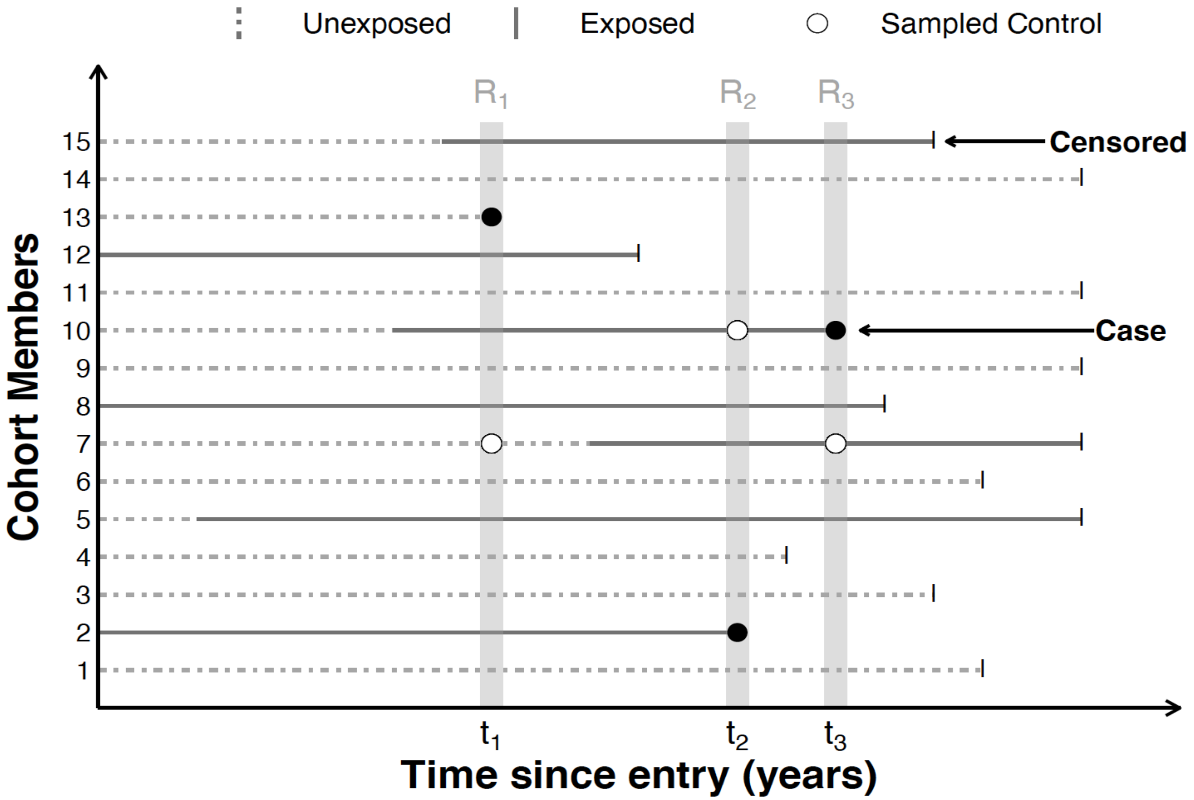
3(b): Power associated with Control-case ratio for a giver number of cases using traditional case control design



Exposure density sampling
(for a time-dependent exposure)

Quick look at Cox model

Time-dependent exposure



NCC sampling and conditional logistic regression:
HR for exposed (yes/no, level, duration) vs. unexposed

Cohort approach: data gathered retrospectively

Example: association of length of hospital stay with exposures during the hospitalisation (e.g. nosocomial infection)*.

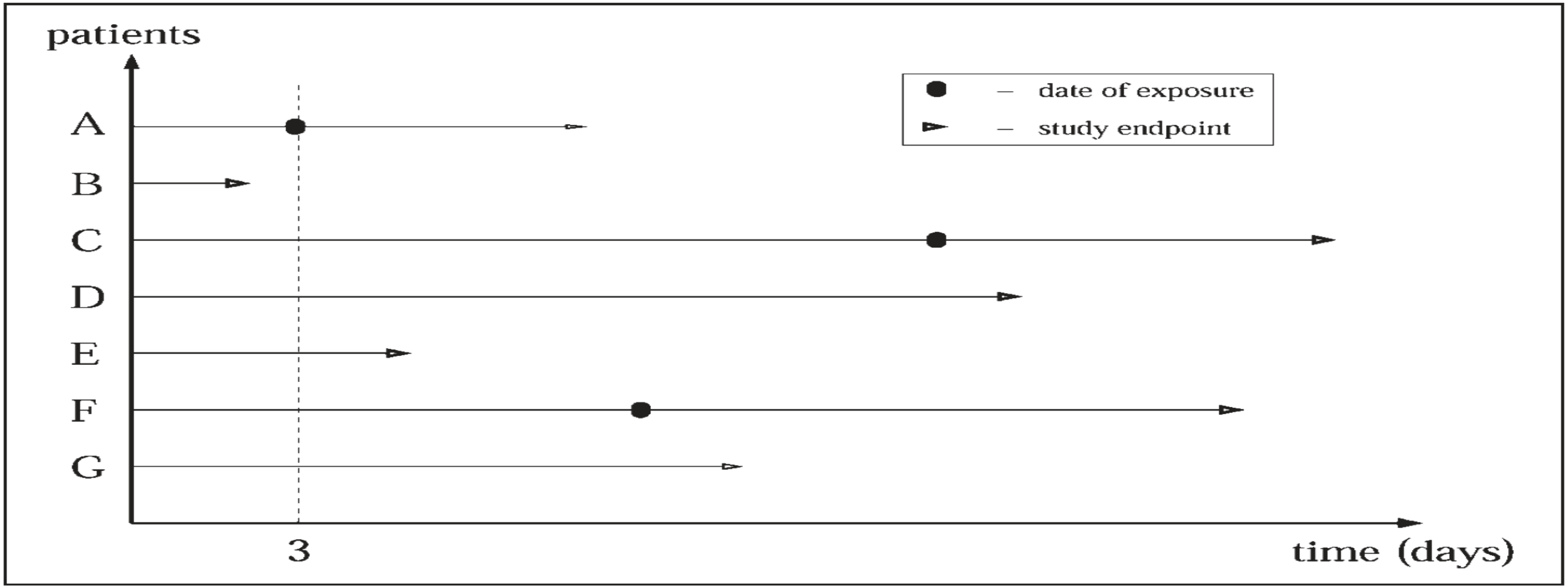


Fig. 1 Example for risk set sampling: For patient A, who gets exposed at day 3, patients C–G are suitable partners by using exposure density sampling whereas only patients D, E and G can be selected using matching for time to exposure.

* *M Wolkewitz, J.Beyersmann, P Gastmeier, M.Schumacher. Meth Inf Med, 2009*

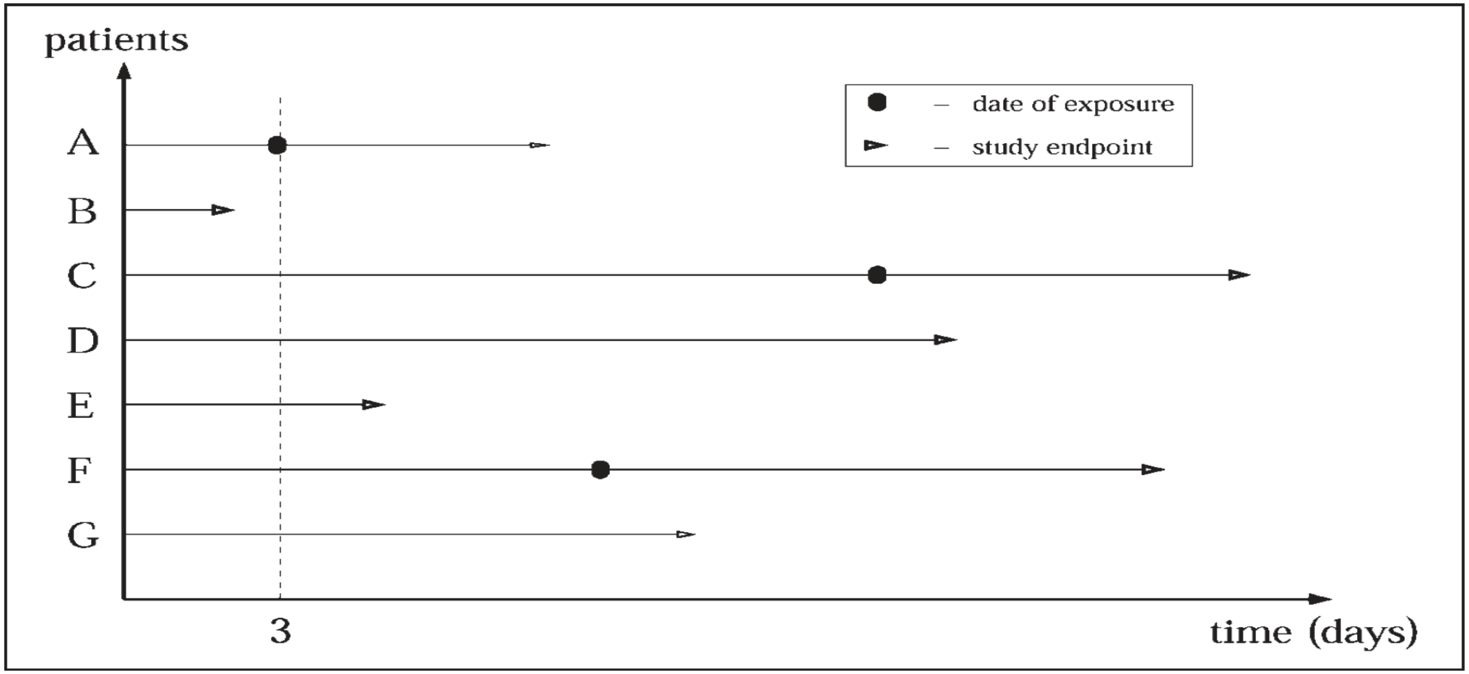


Fig. 1 Example for risk set sampling: For patient A, who gets exposed at day 3, patients C–G are suitable partners by using exposure density sampling whereas only patients D, E and G can be selected using matching for time to exposure.

"Matching on time to exposure":

- For each exposed person, match unexposed persons who have been in hospital at least as long as the time-to-exposure
- *selected from patients who remained unexposed throughout*
- Cox regression (time zero = exposure/matched)

Common in hospital epidemiology

Exposure density sampling

Same principle as incidence density sampling

unexposed can later become exposed

Removes the “time-dependent bias” (also called “survival bias”)

Standard Cox regression for time-dependent covariates (with robust variance)

Potential applications:
Discontinuation of treatment in a cohort
Outcome after (waiting for) treatment in clinical cohort

Countermatching

Matching

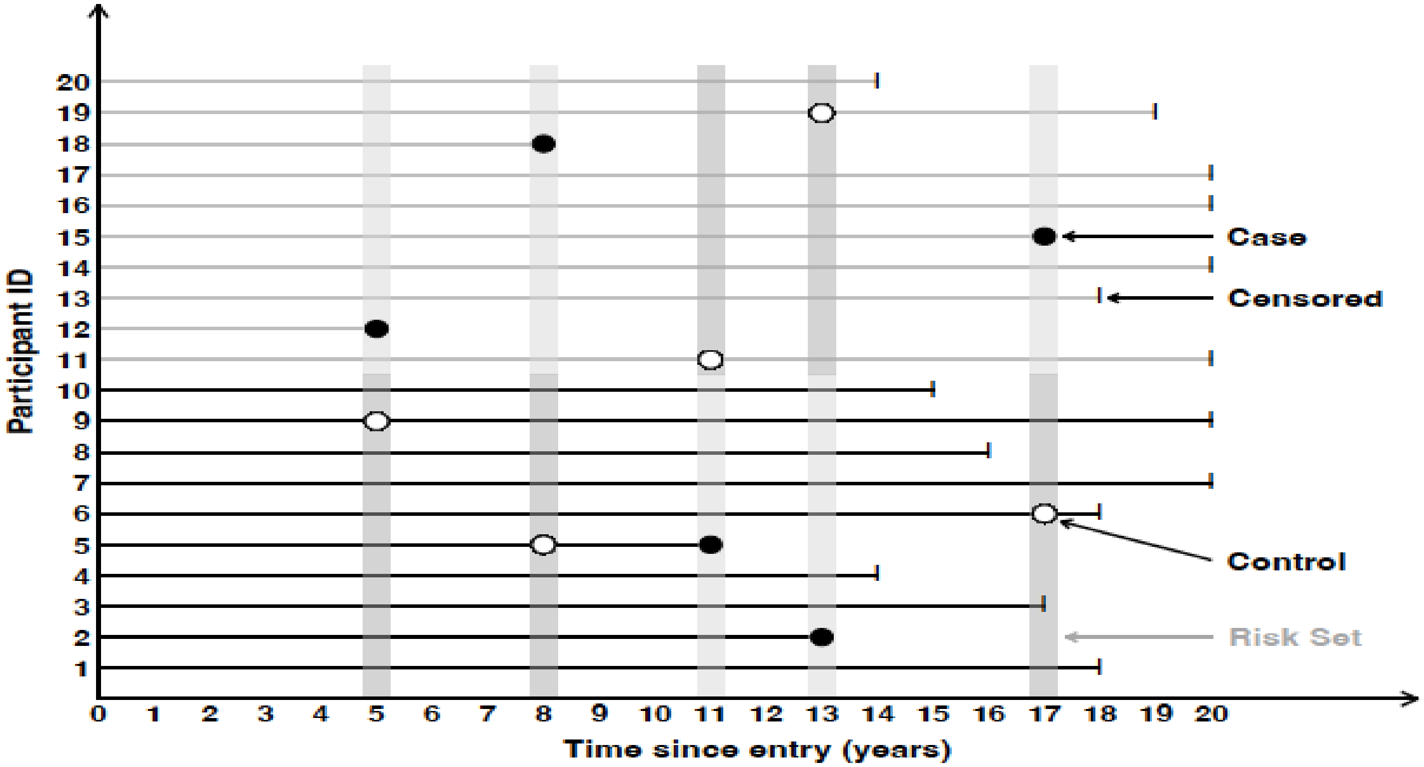
- Purpose: to make cases and controls as similar as possible
- Match on variables not of interest (confounders)
- The effect of the matching factor cannot be estimated by standard methods

Countermatching*

- Purpose: to make cases and controls as different as possible
- **Countermatch on exposure or surrogate of exposure**
- Wider range of exposure improves precision

Langholz and Clayton, Env. Health Persp, 1994

1:1 countermatching



Estimates obtained from weighted conditional likelihood

need risk set sizes in sampling strata in study base

”Counterintuitive matching?

Cologne, commentary in Epidemiology 1997

….. still not widespread, despite:

good efficiency

ability to get estimates from standard software (weights, offset)

Custom R commands at

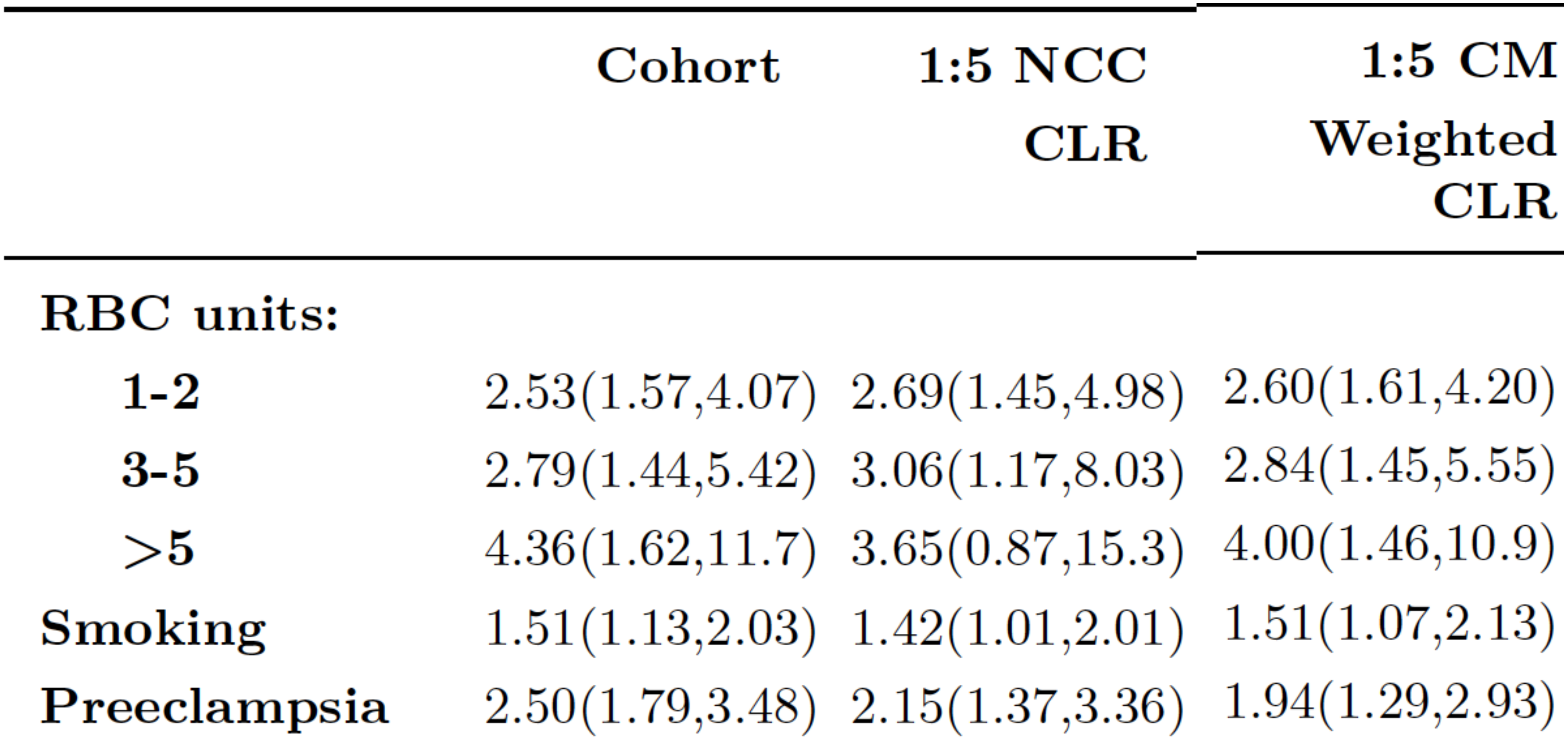
https://github.com/nyilin/SamplingDesignTools

data preparation not difficult

Return to transfusions and post-partum VTE

966,070 deliveries, 472 cases of VTE within 6 weeks of delivery

1:5 NCC study had 84% of sets concordant for exposure!



	Cohort	1:5 NCC CLR	1:5 CM Weighted CLR
RBC units:			
1-2	2.53(1.57,4.07)	2.69(1.45,4.98)	2.60(1.61,4.20)
3-5	2.79(1.44,5.42)	3.06(1.17,8.03)	2.84(1.45,5.55)
>5	4.36(1.62,11.7)	3.65(0.87,15.3)	4.00(1.46,10.9)
Smoking	1.51(1.13,2.03)	1.42(1.01,2.01)	1.51(1.07,2.13)
Preeclampsia	2.50(1.79,3.48)	2.15(1.37,3.36)	1.94(1.29,2.93)

Summary

- Many standard epidemiology designs can be made more efficient by exploiting **exposure-dependent** sampling.

- Benefit in cost-efficiency for investment in design/analysis

- Standard methods (some using re-weighting) provide valid cohort estimates

- Greatest potential for savings where exposure information is costly (e.g. molecular/genetic studies)