The Case-Cohort design: What it is and how it can be used in register-based research

Anna L.V. Johansson

anna.johansson@ki.se

Collaborators: Paul C. Lambert, Therese M-L. Andersson, Paul W. Dickman

Stata Users Group Meeting, Oslo 2016-09-13

Motivation

- In epidemiology, the cohort design is a standard study design, which is characterised by
  - A disease-free population at start of follow-up
  - Which is followed until outcome of interest (disease) or censoring (lost-to-follow-up)
- In register-based epidemiology, national population registers are often used and linked together (using the PIN)
  - Register-based cohorts can be nation-wide
  - Millions of individuals can be followed for decades for an outcome
- The analysis of such nation-wide cohorts can be computationally challenging

Motivation

- In situations when we do not want to (or are unable to) use a full cohort, we often consider a case-control design (to reduce the comparison group)
  - **Traditionally:** Expensive data collection of exposures , e.g. biomarker samples, genotyping , medical records, or questionnaires
  - **NEW**: Reduce data sizes for computational efficiency, e.g. complex modelling, correlated data, multiple timescales
- Today, we have a lot of computational power available
  - But, there are situations when clever subsampling can create more manageable analytical datasets so that a complex model can run faster and even locally on a computer
  - As a statistician doing lots of modelling, I like being able to do that!

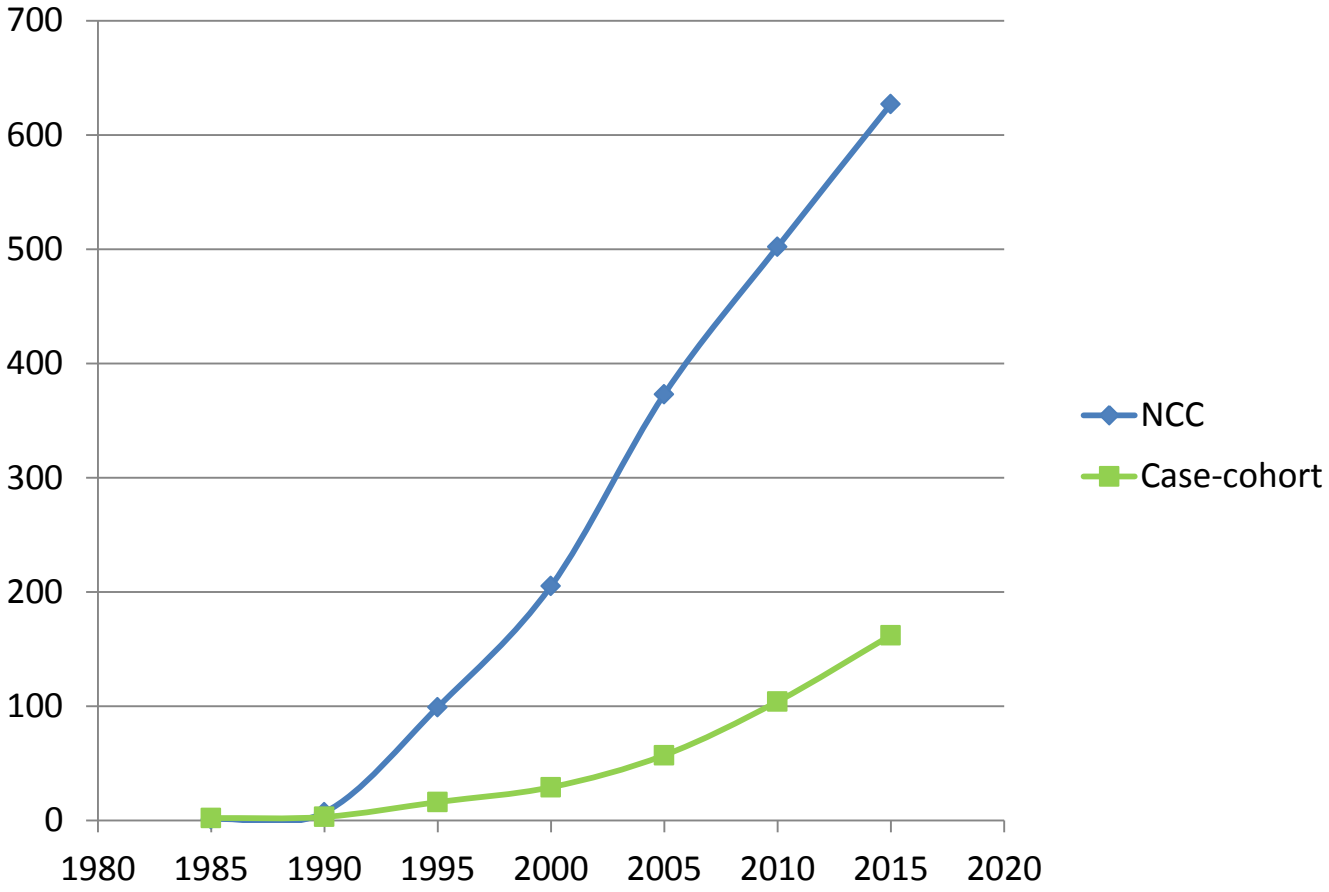
Case-control designs

- **Nested case-control design (NCC**) is an option
  - With appropriate sampling and analysis, the OR estimates the HR in the full cohort

- **Case-cohort design** is another option
  - With appropriate sampling and analysis, the HR estimates the HR in the full cohort
  - In a case-cohort study you can also estimate e.g. rates, rate differences, risks
  - That is an advantage of the case-cohort design over the NCC, where you typically only estimate relative measures (HR) and not absolute measures (hazard rates or risks)

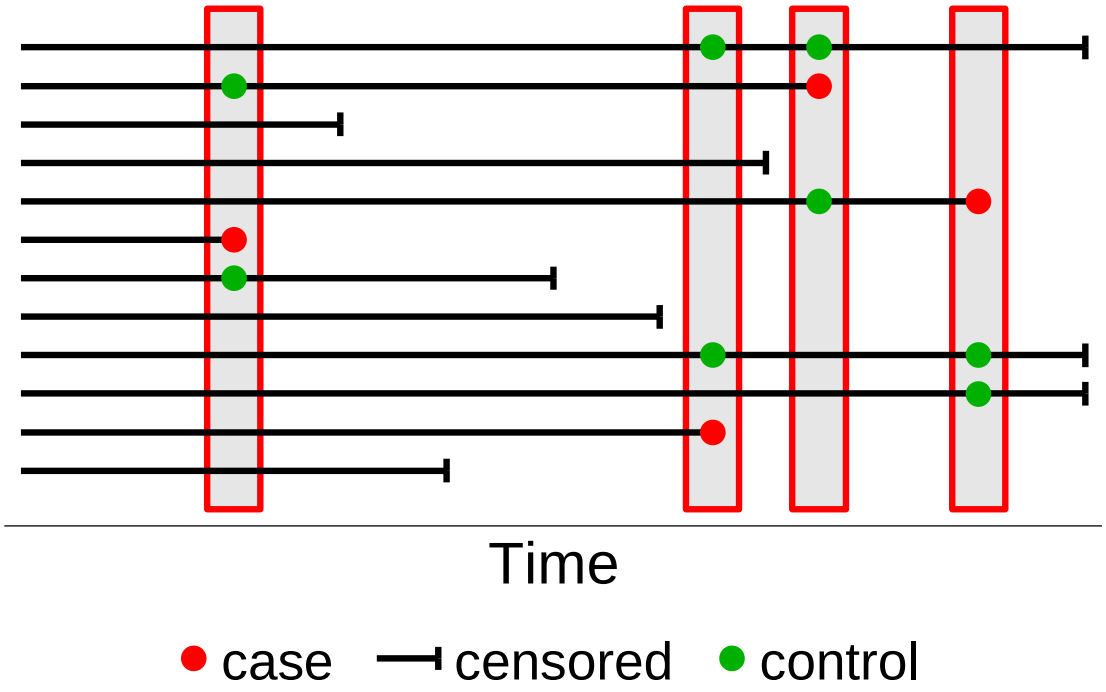
- Case-cohort studies are much less common than NCC studies in literature
  - Design and analysis is thought to be complex – not true anymore!
  - Aim of this talk is to show that case-cohort studies can be easily performed and analysed

References to nested case-control and case-cohort in Web of Science



Nested Case-Control design

Nested Case-Control design (NCC)



Controls are time-matched to cases.
I.e. controls can only be used for <u>one</u> outcome.

Nested Case-Control design (NCC)

- **Sampling of the NCC:**
  - Study base is some large cohort.
  - Select all those who become cases.
  - Sampling of controls (incidence density sampling):
    - Select controls randomly from those still at risk at time of the case ("riskset")
    - Usually 1 to 5 controls per case (>5 controls only improves efficiency minorly)
    - Controls are **time-matched** to cases. (1) Persons can be controls more than once, (2) A person selected as control may later become a case.

- Often involves additional matching on confounders.
- Analysis using conditional logistic regression, conditioning on riskset (and matching strata)
- The odds ratio (OR) estimates the underlying HR in the cohort

- *Originally proposed by Thomas (1977) and developed by Prentice and Breslow (1978)*

Nested Case-Control design (NCC)

- 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 riskset then controls can be re-used.*

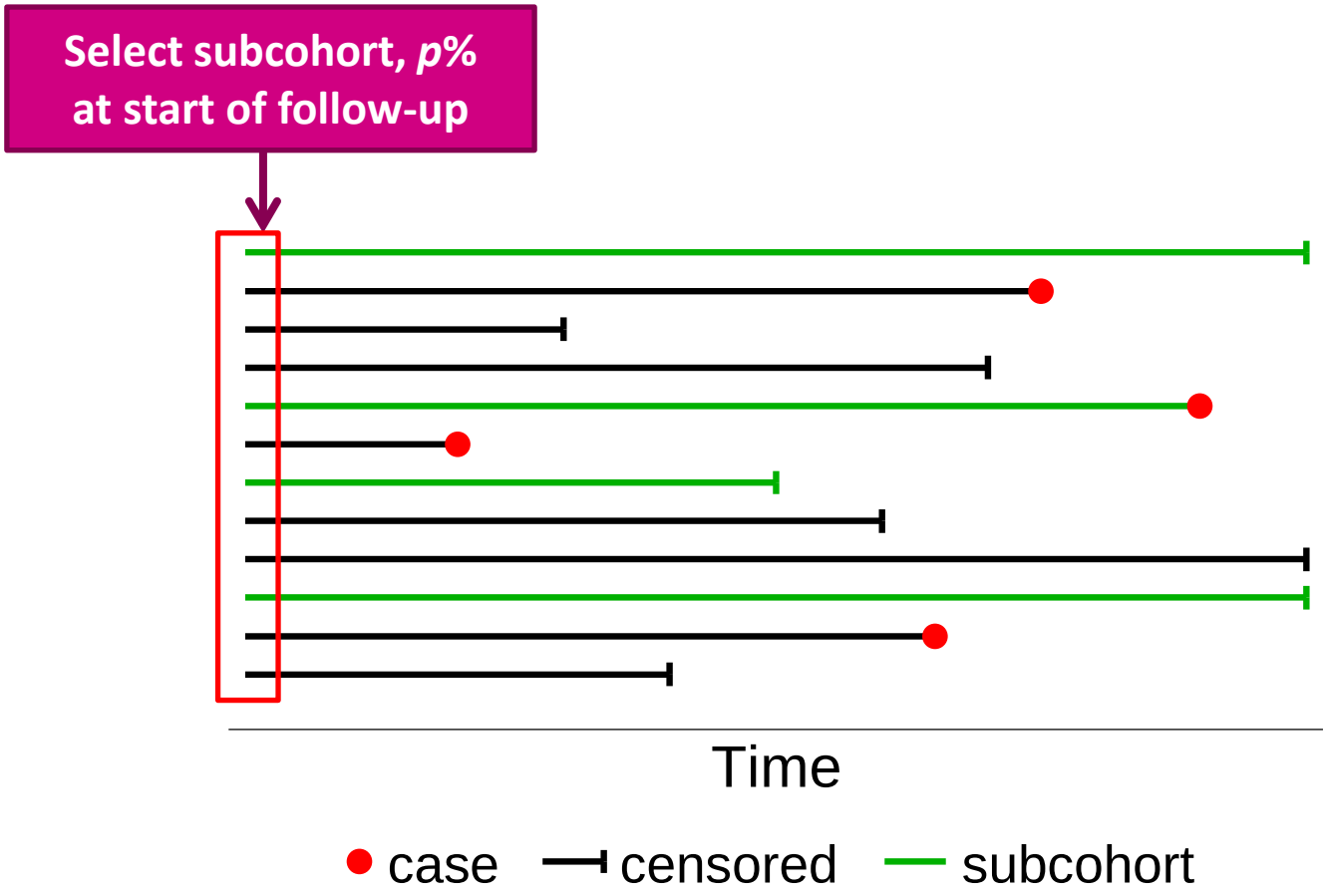
- Limitation 2:
  - We can only estimate HRs, relative rates
  - We cannot estimate rates or risks, since we do not know the underlying persontime at risk (sampling has distorted this information by selecting a fix number of controls from each riskset)
  - *If we know the size of risksets and sampling fractions in each riskset, then it is possible to estimate rates (Langholz, Borgan 1997 and others). Not trivial, especially if there are time-dependent effects.*

Case-cohort design

Case-cohort design

- We start with a cohort study….

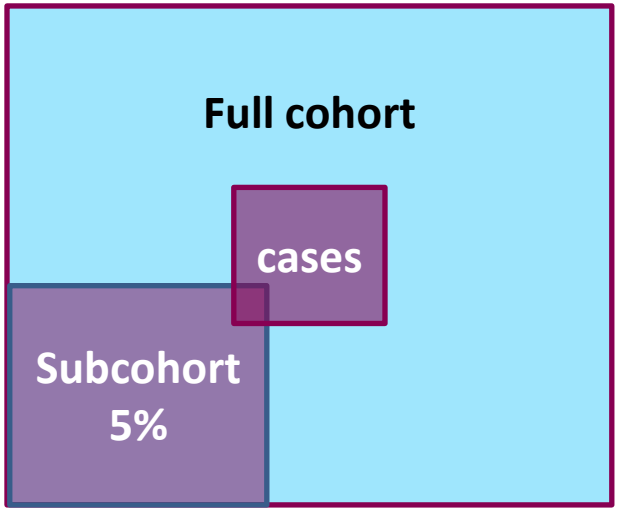
Case-Cohort design



Subcohort is not time-matched to cases.
I.e. controls can be used for many outcomes.

Case-Cohort design

- **Sampling of case-cohort:**
  - From the cohort, select a subcohort of individuals at start of follow-up.
  - The subcohort will include some cases.
  - Also include all cases that occur outside the subcohort during follow-up.
  - Final sample consists of subcohort + cases outside subcohort.
- HR can be estimated, but also hazard rates.
  - Information about population at risk is maintained via the sampling fraction
- Same subcohort can be used for several diseases (outcomes).



Case-Cohort design

- 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 Oslo!)

Analysis of Case-Cohort design

Analysis of Case-Cohort design

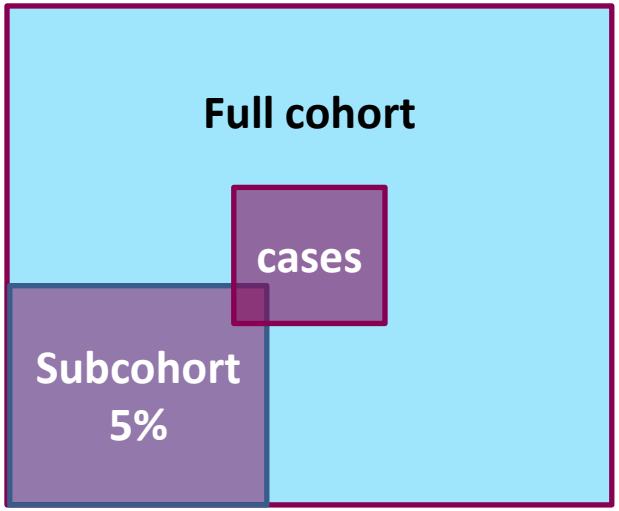
- You need to keep track of persons inside/outside subcohort, and cases/noncases

	In subcohort		
	No (outside)	Yes (inside)	Total
Non-case	M^0	M^s	M
Case	D^0	D^s	D
Total	$\mathbf{T^0}$	$\mathbf{T^s}$	**T**

Sampling fraction: $p = \frac{T^s}{T} = \mathbf{0.05}$

Sampling fraction non-cases: $p_M = \frac{M^s}{M} \approx \mathbf{0.05} = \frac{T^s}{T}$

Sampling fraction cases: $p_D = \frac{D^0 + Ds}{D} = \mathbf{1}$



Analysis of Case-Cohort design

- The analysis of case-cohort studies is thought to be complicated.
  - This is not true anymore.
- *Design and methodology was proposed by Prentice 1986.*
  - *Previous work by Kupper et al (1975) and Miettinen (1982)*

- The analysis includes (in addition to a standard cohort analysis)
  - **Weighting:** Due to oversampling of cases, the analysis must be weighted to produce unbiased estimates of the full cohort.
  - **Adjustment of variance:** Because the same control population is upweighted and used repeatedly over time, the variation is too small, the variance must be adjusted (robust std err, sandwich estimator).

- The literature has focused on modifications of the partial likelihood in the Cox model.
  - **Parametric models can also be used** (Moger et al, 2008), e.g. Poisson regression and Flexible Parametric survival Models (FPM), which are useful with multiple timescales and if interest is in estimating (absolute) hazard rates.

Weighted likelihood approach

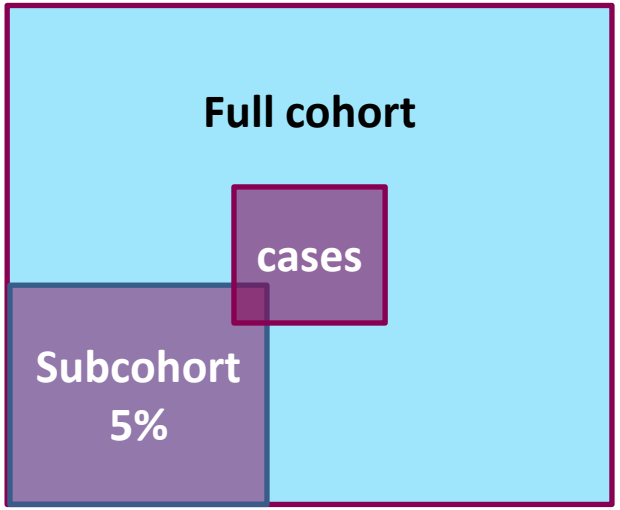
- Several types of weighting schemes have been proposed
  - Good overview in Kulathinal et al (2007); several papers compare different types of weights, not all weights give inference for the full cohort

- Weights based on inverse probability weighting (IPW):
  - Gives inference for the full cohort!
  - Weighted likelihood using "**Borgan II weights"** [Borgan et al, 2000]
    - For cases: w=1
    - For non-cases: w=1/p_M *(one over the sampling fraction of non-cases)*
  - All non-cases are upweighted so that each sampled non-case represents 1/p_M non-cases in the full cohort (if p_M=5% then 1/p_M=20)

- **Weighted likelihood** approach: Cox model or parametric model
  - A weighted likelihood is a ***pseudo-likelihood***, can be used for estimating parameters and CIs, but LR tests are not valid (Wald tests are ok)
  - Need to correct standard errors (upweighting the same subcohort individuals, too little variation), robust std err (sandwich estimator)

How to in Stata

- For the purpose of this presentation, I want to compare an analysis of the full cohort to a case-cohort sample



- Swedish women born 1948-1952 (N=323,850)
  - Breast cancers occurring in ages 25-50 years.
- **Sampling of case-cohort** design:
  - A subcohort of 5% was randomly drawn.
  - All breast cancer cases occurring outside the subcohort were included.
- Modelled **educational level** (high vs low) as the only covariate.
  - Compare: Full cohort and Case-cohort
  - Compare: Cox model and Flexible Parametric model

How to in Stata: Create the case-cohort sample

```
. set seed 339487731  // makes sampling reproducible
. gen u = runiform()   // assign random number to all obs
. gen subcoh = u < 0.05 // generate dummy subcohort
. tab case subcoh
```

case	subcoh 0	1	Total
0	302,939	15,990	318,929
1	4,692	229	4,921
Total	307,631	16,219	323,850

Sampling fraction non-cases:

$$p_M = \frac{15,990}{318,929} = 0.050137$$

Sampling fraction, total:

$$p = \frac{16,219}{323,850} = 0.050082$$

Full cohort: n= 323,850

Case-cohort: n= 20,911 (i.e. 15,990 + 4,692+229)

How to in Stata: Define the cohort

```
. stset exitdate, fail(bc_event==1) enter(time date_age25)
                  exit(time date_age50) ///
                  origin(mother_birthdate) ///
                  scale(365.24) id(lopnrmor)

                id:  lopnrmor
     failure event:  bc_event == 1
obs. time interval:  (exitdate[_n-1], exitdate]
 enter on or after:  time date_age25
 exit on or before:  time date_age50
    t for analysis:  (time-origin)/365.24
            origin:  time mother_birthdate
------------------------------------------------------------------------------
   324699  total obs.
      352  ignored because never entered
      497  obs. end on or before enter()
------------------------------------------------------------------------------
   323850  obs. remaining, representing
   323850  subjects
     4921  failures in single failure-per-subject data
  8011716  total analysis time at risk, at risk from t =         0
                            earliest observed entry t =        25
                                 last observed exit t =  50.00274
```

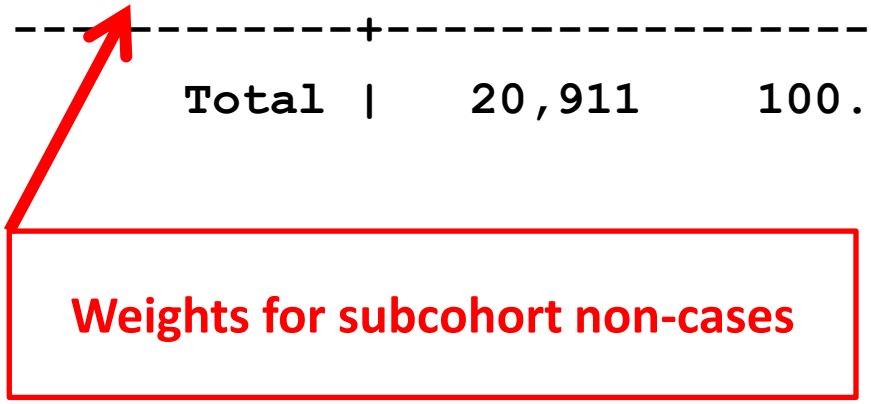
```
. gen case=_d
```

NOTE: IMPORTANT! Define case based on _d

How to in Stata

```
// Generate Borgan II weights
. gen wt = 1 if case==1
. replace wt = 1 / (15,990/318,929) if case==0 & subcoh==1
```

wt	Freq.	Percent	Cum.
1	4,921	23.53	23.53
19.94553	15,990	76.47	100.00
Total	20,911	100.00	



How to in Stata: Weighted models

```
/* STSET using pweights option*/
. stset exitdate [pw=wt], fail(bc_event==1) ///
                          enter(time date_age25) ///
                          exit(time date_age50) ///
                          origin(mother_birthdate) ///
                          scale(365.24) id(lopnrmor)

/* Cox model for case-cohort - Borgan II*/
. stcox educ2, vce(robust)

/* FPM model for case-cohort - Borgan II */
. stpm2 educ2, scale(h) df(5) eform vce(robust) nolog
```

Results Breast Cancer

Table: Comparing full cohort to case-cohort (5%). HR for High vs. Low Education.

		Cox Model	Flexible Parametric Model
Full cohort	HR	0.8363	0.8363
	β	-0.1787	-0.1787
	Std Err	0.0318	0.0318
Case-cohort (Borgan II)	HR	0.8270	0.8270
	β	-0.1900	-0.1900
	Std Err*	0.0358	0.0358

* vce(robust)

Full cohort n=323,850, cases n=4,921

Case-cohort n=20,911, cases n=4,921

Should be similar.
Sampling variation may cause HRs to differ.

Results Breast Cancer

Table: Comparing full cohort to case-cohort (5%). HR for High vs. Low Education.

		Cox Model	Flexible Parametric Model
Full cohort	HR	0.8363	0.8363
	β Std Err	-0.1787 0.0318	-0.1787 0.0318
Case-cohort (Borgan II)	HR	0.8270	0.8270
	β Std Err*	-0.1900 0.0358	-0.1900 0.0358

* vce(robust)

Full cohort n=323,850, cases n=4,921

Case-cohort n=20,911, cases n=4,921

The additional error for case-cohort is very small in comparison to the gain in dataset reduction.

Results Breast Cancer

Table: Comparing full cohort to case-cohort (5%). HR for High vs. Low Education.

		Cox Model	Flexible Parametric Model
Full cohort	HR	0.8363	0.8363
	β Std Err	-0.1787 0.0318	-0.1787 0.0318
Case-cohort (Borgan II)	HR	0.8270	0.8270
	β Std Err*	-0.1900 0.0358	-0.1900 0.0358

* vce(robust)

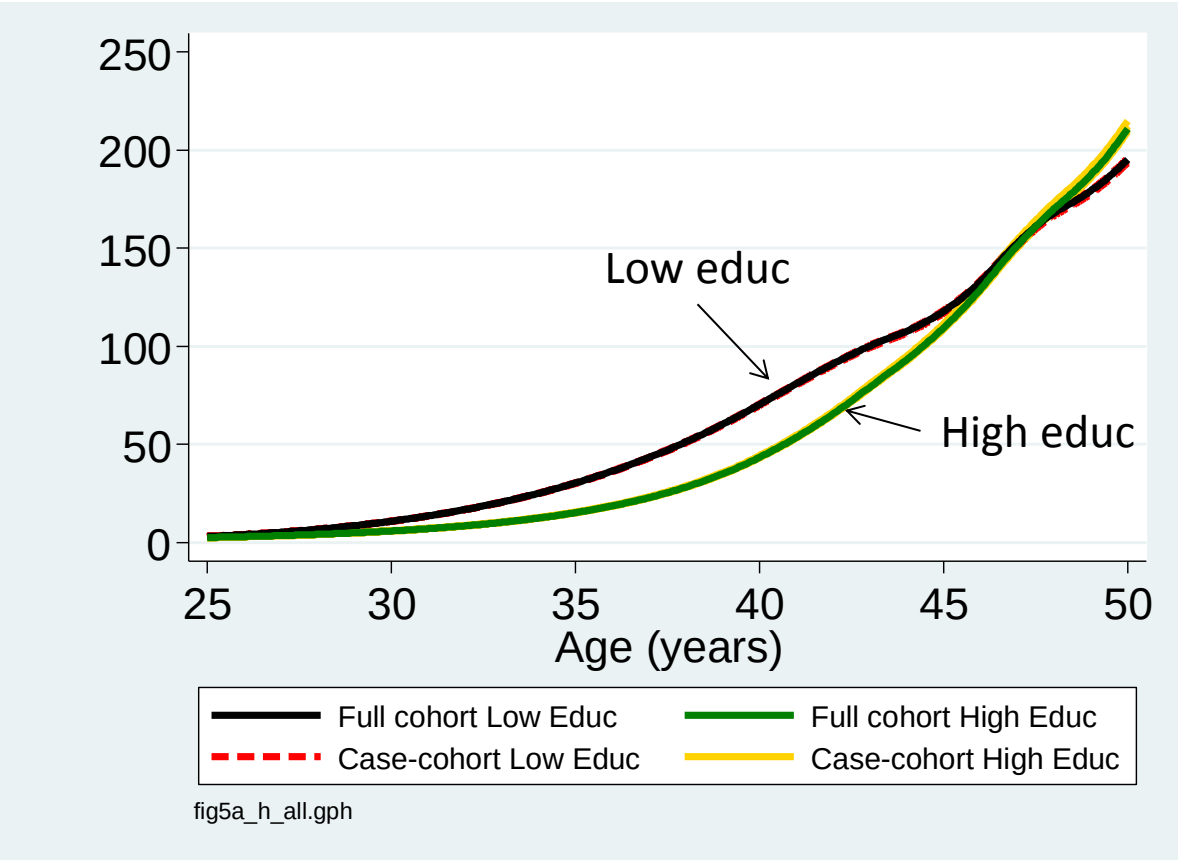
Full cohort n=323,850, cases n=4,921

Case-cohort n=20,911, cases n=4,921

Results from Cox and FPM are similar!

Incidence rates: Hazard by education level

- Time-varying incidence rates (allowing for non-proportional hazards)
- Small variation in results between case-cohort samples and full cohort



Stpm2, Borgan II weights, non-prop haz. model

In summary

In summary

- The design and analysis of case-cohort studies is straight-forward!
  - Pweight option is great for this in Stata!

- Situations when the case-cohort design is useful
  - **Traditionally**: Expensive data collection on exposures or multiple endpoints
  - **New**: Reduce analytical dataset for computational efficiency
  - Interest is in absolute measures (rates, rate diff's, risks), not just relative rates

My study: Pregnancy and BC, case-cohort, multiple timescales

Breast Cancer Res Treat (2015) 151:209–217
DOI 10.1007/s10549-015-3369-4

EPIDEMIOLOGY

Family history and risk of pregnancy-associated breast cancer (PABC)

Anna L. V. Johansson[1] · Therese M.-L. Andersson[1] · Chung-Cheng Hsieh[2] · Sven Cnattingius[3] · Paul W. Dickman[1] · Mats Lambe[1,4]

Received: 1 April 2015 / Accepted: 2 April 2015 / Published online: 19 April 2015

Abstract The risk of breast cancer is at least two-fold increased in young women with a family history of breast cancer. Pregnancy has a dual effect on breast cancer risk: a short-

this peak was only present in women with a family history. Our results indicate that women with a family history of breast cancer do not have a different breast cancer risk during and

References

- ***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, Samuelsen)***
- **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;** *teaching notes from 2005* - http://folk.uio.no/osamuels/casecohort4.pdf
- **Borgan, Samuelsen (2003):** A review of cohort sampling designs for Cox's regression model: Potentials in epidemiology. *Norsk Epidemiologi, 13:239-248. 2003*
- **Lambert, Royston (2009);** Further development of flexible parametric models for survival analysis. *Stata Journal 2009.*
- **Cologne et al (2012);** Conventional case–cohort design and analysis for studies of interaction. *International Journal of Epidemiology 2012;1–13*
- **Johansson AL et al (2015)**. Breast Cancer Res Treat 2015; 151: 209-217.
- **Johansson AL et al.** *Analysing case-cohort data using flexible parametric survival models (FPM). In manuscript*

References

Examples of epi studies which have used the case-cohort design:

- **Karvanen et al (2009);** The impact of newly identified loci on coronary heart disease, stroke and total mortality in the MORGAM prospective cohorts. *Genet Epidemiol, 2009*.
- **Luft et al (2015);** Carboxymethyl lysine, and advanced glycation end product, and incident diabetes: a case-cohort analysis of the ARIC study. *Diabetic Medicine 2015*
- **Geybels et al (2014);** Selenoprotein gene variants, toenail selenium levels, and risk for advanced prostate cancer. *JNCI, 2014*