Nested case-control and case-subcohort studies

Christiana Kartsonaki

Nuffield Department of Population Health

7 July 2017

Introduction

Studies within prospective cohort studies

- Nested case-control studies
- Case-subcohort studies

Introduction

Studies within prospective cohort studies

- Nested case-control studies
- Case-subcohort studies

Considerations for their design and analysis.

Studies nested within prospective cohort studies

Existing prospective cohort study, want to measure additional exposure(s) and assess their associations with outcome(s).

Studies nested within prospective cohort studies

Existing prospective cohort study, want to measure additional exposure(s) and assess their associations with outcome(s).

A problem might be that it is not feasible to collect data on the additional exposures on all individuals, especially if the original cohort is large, due to cost, time, not wanting to use up existing biological samples.

Even if it is feasible, it may not be necessary.

Nested case-control studies

Sampling schemes for nested case-control studies:

- Disregarding time – *cumulative case-control sampling*
  Most appropriate for short duration of observation (follow-up)

Nested case-control studies

Sampling schemes for nested case-control studies:

- Disregarding time – *cumulative case-control sampling*
  Most appropriate for short duration of observation (follow-up)

- Using time – *(incidence) density sampling*
  Controls are selected when a case developed the outcome (matching on time)

Nested case-control studies

Sampling schemes for nested case-control studies:

- Disregarding time – *cumulative case-control sampling*
  Most appropriate for short duration of observation (follow-up)

- Using time – *(incidence) density sampling*
  Controls are selected when a case developed the outcome (matching on time)

The underlying cohort need not be an assembled cohort of individuals participating in a study but may be any well defined sampling frame, such as a country-wide health registry.

Nested case-control studies

Disregarding time (cumulative case-control sampling)

All individuals in the population were in principle observable for the whole of the follow-up period during which cases arose.

Nested case-control studies

Disregarding time (cumulative case-control sampling)

All individuals in the population were in principle observable for the whole of the follow-up period during which cases arose.

Sometimes unrealistic, especially when the time period over which cases are ascertained is long.

- how to treat individuals who leave/join the study population
- what to do about exposures and other explanatory variables that may change appreciably over time

Nested case-control studies

Disregarding time (cumulative case-control sampling)

All individuals in the population were in principle observable for the whole of the follow-up period during which cases arose.

Sometimes unrealistic, especially when the time period over which cases are ascertained is long.

- how to treat individuals who leave/join the study population
- what to do about exposures and other explanatory variables that may change appreciably over time

Why not use as controls a group of individuals who survive to the end of the observation period?

- if the observation period had been shorter then, for cases occurring early in the observation time period, the pool of potential controls would be different from that if the observation period had been longer
- concerns whether individuals who survive over a long time period without being censored for some reason intrinsically different from the rest of the underlying population of non-cases

Nested case-control studies

Using time

We will from now on use the term nested case-control studies to refer to the design that takes into account event times in the sampling of controls.

Nested case-control studies

Using time

We will from now on use the term nested case-control studies to refer to the design that takes into account event times in the sampling of controls.

One or more controls are selected for each case from the **risk set** at the time at which the case arose.

Risk set: the set of all individuals still in the study at that time (still eligible to experience and have the event observed at that time, that is, still event-free and not censored).

Nested case-control studies

Using time

We will from now on use the term nested case-control studies to refer to the design that takes into account event times in the sampling of controls.

One or more controls are selected for each case from the **risk set** at the time at which the case arose.

Risk set: the set of all individuals still in the study at that time (still eligible to experience and have the event observed at that time, that is, still event-free and not censored).

Individuals can be sampled as controls for more than one case and individuals sampled as controls may subsequently become cases.

Nested case-control studies

Risk sets differ depending on whether time since entry into the study or age is used as the underlying time scale

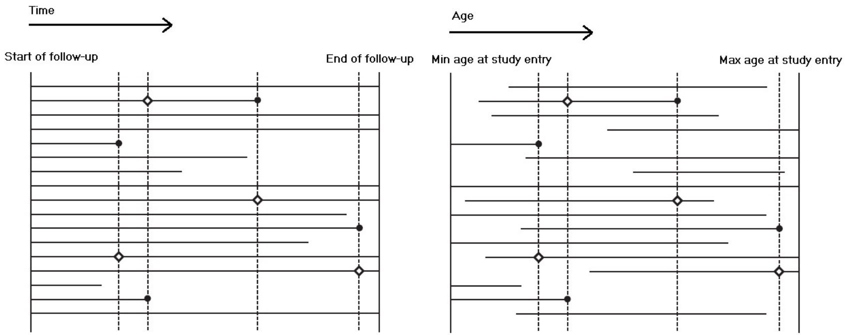

Figure: Sampling a nested case-control study with one control per case, using time since study recruitment or age as the timescale. The solid lines represent the time period over which individuals are observed. Cases occur over the course of follow-up (•), individuals may leave the population or may survive to the end of the follow-up period or maximum age of observation. The dotted lines pass through members of the risk set at each event time. One control (○) is selected for each case from its risk set. Adapted from Keogh and Cox (2014).

Nested case-control studies

- Can have one or more controls per case

Nested case-control studies

- Can have one or more controls per case
- Can match on other variables, but care needed not to overmatch

Nested case-control studies

- Can have one or more controls per case
- Can match on other variables, but care needed not to overmatch
- Analysis is done using a Cox proportional hazards model and a modification to the partial likelihood (weighted PL) used in full-cohort studies, yielding estimates of hazard ratios

Nested case-control studies

- Can have one or more controls per case
- Can match on other variables, but care needed not to overmatch
- Analysis is done using a Cox proportional hazards model and a modification to the partial likelihood (weighted PL) used in full-cohort studies, yielding estimates of hazard ratios
- Same as conditional logistic regression under this simple sampling scheme

Nested case-control studies

- Can have one or more controls per case
- Can match on other variables, but care needed not to overmatch
- Analysis is done using a Cox proportional hazards model and a modification to the partial likelihood (weighted PL) used in full-cohort studies, yielding estimates of hazard ratios
- Same as conditional logistic regression under this simple sampling scheme
- Can estimate absolute risks

Nested case-control studies

- Can have one or more controls per case
- Can match on other variables, but care needed not to overmatch
- Analysis is done using a Cox proportional hazards model and a modification to the partial likelihood (weighted PL) used in full-cohort studies, yielding estimates of hazard ratios
- Same as conditional logistic regression under this simple sampling scheme
- Can estimate absolute risks
- Can accommodate time-varying exposures and recurrent events

Nested case-control studies

- Can have one or more controls per case
- Can match on other variables, but care needed not to overmatch
- Analysis is done using a Cox proportional hazards model and a modification to the partial likelihood (weighted PL) used in full-cohort studies, yielding estimates of hazard ratios
- Same as conditional logistic regression under this simple sampling scheme
- Can estimate absolute risks
- Can accommodate time-varying exposures and recurrent events
- Other survival models are possible (e.g. fully parametric models, additive hazards models)

Example

Pai et al. (2004)

Nurses' Health Study and Health Professionals Follow-up Study

Large US prospective cohort studies

Blood samples collected a few years after recruitment

Example

Pai et al. (2004)

Nurses' Health Study and Health Professionals Follow-up Study

Large US prospective cohort studies

Blood samples collected a few years after recruitment

Associations between inflammatory markers and the risk of coronary heart disease

Example

Pai et al. (2004)

Of the 121,700 Nurses' Health Study participants, 32,826 provided a blood sample and of 51,529 Health Professionals Study participants 18,225 did so.

Example

Pai et al. (2004)

Of the 121,700 Nurses' Health Study participants, 32,826 provided a blood sample and of 51,529 Health Professionals Study participants 18,225 did so.

249 female cases who had a nonfatal myocardial infarction or fatal coronary heart disease between the date of the blood sample and June 1998

266 male cases were those who had such an event between the date of the blood sample and 2000

Example

Pai et al. (2004)

Of the 121,700 Nurses' Health Study participants, 32,826 provided a blood sample and of 51,529 Health Professionals Study participants 18,225 did so.

249 female cases who had a nonfatal myocardial infarction or fatal coronary heart disease between the date of the blood sample and June 1998

266 male cases were those who had such an event between the date of the blood sample and 2000

At each case's event time 2 controls were randomly sampled from the risk set, excluding the case, matched to the case on age, smoking status, date of blood sample and, in the male study only, fasting status at the time of the blood sample.

In R

Function ccwc (by David Clayton) in package Epi (Carstensen et al., 2017) samples controls for a nested case-control study

Function clogit (by Thomas Lumley) in package survival (Therneau, 2015) fits a conditional logistic regression model

In R

Generating a nested case-control study

```
ccwc(entry = 0, exit, fail, origin = 0, controls
    = 1, match = list(), include = list(), data
   = NULL, silent = FALSE)
```

`entry`: Time of entry to follow-up
`exit`: Time of exit from follow-up
`fail`: Status on exit (1 = fail, 0 = censored)
`origin`: Origin of analysis time scale
`controls`: The number of controls to be selected for each case
`match`: List of categorical variables on which to match cases and controls
`include`: List of other variables to be carried across into the case-control study
`data`: Data frame in which to look for input variables
`silent`: If FALSE, echos a `.` to the screen for each case-control set created; otherwise produces no output.

In R

Fitting a conditional logistic regression model

```
clogit(formula, data, weights, subset, na.action
    , method = c("exact", "approximate", "efron",
     "breslow"), ...)
```

`formula`: Model formula

`data`: data frame

`weights`: optional, names the variable containing case weights

`subset`: optional, subset the data

`na.action`: optional na.action argument. By default the global option na.action is used.

`method`: use the correct (exact) calculation in the conditional likelihood or one of the approximations

`...`: optional arguments, which will be passed to `coxph.control`

Nested case-control studies

Can we do better?

- in some cases efficiency can be increased using techniques such as countermatching and quota sampling (also backwards and forwards use of controls)
- in some cases multiple imputation can be used to utilise information on the full cohort

Nested case-control studies

Can we do better?

- in some cases efficiency can be increased using techniques such as countermatching and quota sampling (also backwards and forwards use of controls)
- in some cases multiple imputation can be used to utilise information on the full cohort

What if we want to study another outcome in the same study?

- methods for re-using controls for other outcomes exist
- however for multiple outcomes a more natural choice is to use a case-subcohort design

Case-subcohort studies

Also called *case-cohort* studies

- Disregarding times – *case-base sampling* (or *hybrid epidemiological design*)
- Using times

Case-subcohort studies

Also called *case-cohort* studies

- Disregarding times – *case-base sampling* (or *hybrid epidemiological design*)
- Using times

The primary feature of a case-subcohort study is the **subcohort**, which is a random sample from the cohort at study baseline, selected ignoring any information obtained during follow-up, and which serves as the set of potential controls for all cases.

The study comprises the subcohort plus all additional cases, that is, those not in the subcohort.

Case-subcohort studies

Also called *case-cohort* studies

- Disregarding times – *case-base sampling* (or *hybrid epidemiological design*)
- Using times

The primary feature of a case-subcohort study is the **subcohort**, which is a random sample from the cohort at study baseline, selected ignoring any information obtained during follow-up, and which serves as the set of potential controls for all cases.

The study comprises the subcohort plus all additional cases, that is, those not in the subcohort.

Case-subcohort studies

Also called *case-cohort* studies

- Disregarding times – *case-base sampling* (or *hybrid epidemiological design*)
- Using times

The primary feature of a case-subcohort study is the **subcohort**, which is a random sample from the cohort at study baseline, selected ignoring any information obtained during follow-up, and which serves as the set of potential controls for all cases.

The study comprises the subcohort plus all additional cases, that is, those not in the subcohort.

Case-subcohort studies

Disregarding time

- Time period, preferably short, during which cases are observed
- Assumes that individuals who do not become cases are observed for the entire time period
- Odds ratios and risk ratios, logistic regression

Case-subcohort studies

Using time

From now on we consider the analysis using time.

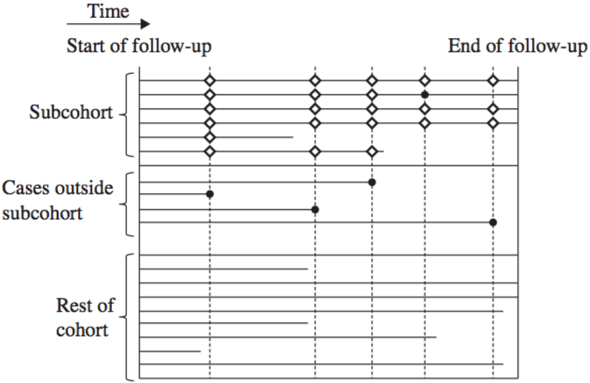

Figure: A case-subcohort study using event times. The horizontal lines show individual follow-up. The cases are indicated by • and the non-cases used in the comparison set at each failure time are indicated by ○. From Keogh and Cox (2014).

Case-subcohort studies

Using time

- The same subcohort can be used for comparisons with different sets of cases

Case-subcohort studies

Using time

- The same subcohort can be used for comparisons with different sets of cases
- The cases are compared with members of the subcohort who are at risk at their event time, using a pseudo-partial likelihood (also called weighted Cox regression) – estimates of hazard or rate ratios

Case-subcohort studies

Using time

- The same subcohort can be used for comparisons with different sets of cases
- The cases are compared with members of the subcohort who are at risk at their event time, using a pseudo-partial likelihood (also called weighted Cox regression) – estimates of hazard or rate ratios
- Sandwich standard errors (correlations between risk sets)

Case-subcohort studies

Using time

- The same subcohort can be used for comparisons with different sets of cases
- The cases are compared with members of the subcohort who are at risk at their event time, using a pseudo-partial likelihood (also called weighted Cox regression) – estimates of hazard or rate ratios
- Sandwich standard errors (correlations between risk sets)
- Variations of this method exist, differing mainly on how risk sets are formed (how the cases outside the subcohort are treated) and on weighting

Case-subcohort studies

- Can use the subcohort for assessing associations with other measures collected at study baseline

Case-subcohort studies

- Can use the subcohort for assessing associations with other measures collected at study baseline
- Can estimate absolute risks

Case-subcohort studies

- Can use the subcohort for assessing associations with other measures collected at study baseline
- Can estimate absolute risks
- Can accommodate time-varying exposures

Case-subcohort studies

- Can use the subcohort for assessing associations with other measures collected at study baseline
- Can estimate absolute risks
- Can accommodate time-varying exposures
- Other survival models are possible

In R

Function `cch` (by Norman Breslow, modified by Thomas Lumley) in package `survival` (Thernau, 2015) fits weighted Cox regression models for case-subcohort studies

In R

Fitting a weighted Cox regression model to case-subcohort data

```
cch(formula, data = sys.parent(), subcoh, id,
   stratum = NULL, cohort.size, method = c("
   Prentice", "SelfPrentice", "LinYing", "I.
   Borgan", "II.Borgan"), robust = FALSE)
```

formula: A formula object that must have a Surv object as the response. The Surv object must be of type "right", or of type "counting".
subcoh: Vector of indicators for subjects sampled as part of the sub-cohort. Code 1 or TRUE for members of the sub-cohort, 0 or FALSE for others. If data is a data frame then subcoh may be a one-sided formula.
id: Vector of unique identifiers, or formula specifying such a vector.
stratum: A vector of stratum indicators or a formula specifying such a vector

In R

Fitting a weighted Cox regression model to case-subcohort data

```
cch(formula, data = sys.parent(), subcoh, id,
   stratum = NULL, cohort.size, method = c("
   Prentice", "SelfPrentice", "LinYing", "I.
   Borgan", "II.Borgan"), robust = FALSE)
```

`cohort.size`: Vector with size of each stratum in the original cohort from which the subcohort was sampled
`data`: An optional data frame in which to interpret the variables occurring in the formula.
`method`: Three procedures are available. The default method is ”Prentice”, with options for ”SelfPrentice” or ”LinYing”.
`robust`: For ”LinYing” only, if `robust = TRUE`, use design-based standard errors even for phase I

The data argument must not have missing values for any variables in the model.

Case-subcohort studies

Can we do better?

- stratified sampling (for example when some strata are rare in the population, but in some cases restricts reusability), stratified analysis (each case is compared with individuals in the same stratum of the subcohort as the case – hazards proportional within strata) or both
- in some cases so-called 'optimal' weights can be used to utilise information on the full cohort
- in some cases multiple imputation can be used to utilise information on the full cohort

Example

Associations of metabolomics and proteomics with risk of pancreatic cancer and diabetes in the China Kadoorie Biobank

China Kadoorie Biobank (CKB) – prospective study of 0.5M Chinese adults

- blood samples collected at baseline (2004–08)
- linkage to health insurance records, disease and death registries

700 pancreatic cancer incident cases accumulated by 2015, as well as a few thousand diabetes cases

Example

Associations of metabolomics and proteomics with risk of pancreatic cancer and diabetes in the China Kadoorie Biobank

China Kadoorie Biobank (CKB) – prospective study of 0.5M Chinese adults

- blood samples collected at baseline (2004–08)
- linkage to health insurance records, disease and death registries

700 pancreatic cancer incident cases accumulated by 2015, as well as a few thousand diabetes cases

Want to use NMR metabolomics assay and proteomics assay on samples collected at baseline to identify markers associated with

1. pancreatic cancer
2. diabetes

and possibly several other diseases in the future

Example

Associations of metabolomics and proteomics with risk of pancreatic cancer and diabetes in the China Kadoorie Biobank

Can’t afford to do either in the whole cohort

Example

Associations of metabolomics and proteomics with risk of pancreatic cancer and diabetes in the China Kadoorie Biobank

Can't afford to do either in the whole cohort

Case-subcohort study

- all 700 pancreatic cancer cases
- a subset of 1000 diabetes cases
- a subcohort of size 1050 (randomly selected from the cohort at baseline)

Example

Armstrong et al. (1994)

Case-subcohort study to investigate whether aluminium plant workers exposed to coal tar pitch volatiles were at increased risk of death from lung cancer.

Source population: over 16,000 men who had worked for at least one year in an aluminium production plant between 1950 and 1979 in Quebec, Canada.

Example

Armstrong et al. (1994)

Case-subcohort study to investigate whether aluminium plant workers exposed to coal tar pitch volatiles were at increased risk of death from lung cancer.

Source population: over 16,000 men who had worked for at least one year in an aluminium production plant between 1950 and 1979 in Quebec, Canada.

Exposures of interest: two indices of exposure to coal tar pitch volatiles

Example

Armstrong et al. (1994)

Case-subcohort study to investigate whether aluminium plant workers exposed to coal tar pitch volatiles were at increased risk of death from lung cancer.

Source population: over 16,000 men who had worked for at least one year in an aluminium production plant between 1950 and 1979 in Quebec, Canada.

Exposures of interest: two indices of exposure to coal tar pitch volatiles

338 lung cancer deaths observed during follow-up (from 1950 to 1988)

Subcohort: random sample of 1138 men

Example

Armstrong et al. (1994)

Company experts estimated the exposure indices for different job types and across calendar time. These estimates were combined with work histories to estimate cumulative exposures.

Confounding by smoking – smoking histories were obtained from company medical records.

Example

Armstrong et al. (1994)

Company experts estimated the exposure indices for different job types and across calendar time. These estimates were combined with work histories to estimate cumulative exposures.

Confounding by smoking – smoking histories were obtained from company medical records.

Time scale: age

Each case was compared with members of the subcohort still alive at the case's age. This required estimation of the cumulative exposures for all at-risk subcohort members at each case's age at death.

Discussion

- Two broad types of sub-studies within prospective cohort studies: nested case-control and case-subcohort studies
- Careful design to ensure reliable results

References

Armstrong, B., Tremblay, C., Baris, D. and Theriault, G. (1994). Lung cancer mortality and polynuclear aromatic hydrocarbons: a case-cohort study of aluminum production workers in Arvida, Quebec, Canada. *American Journal of Epidemiology*, **139**, 250–262.

Barlow, W. E. (1994). Robust variance estimation for the case-cohort design. *Biometrics*, **50**, 1064–1072.

Barlow, W. E., Ichikawa, L., Rosner, D. and Izumi, S. (1999). Analysis of case-cohort designs. *Journal of Clinical Epidemiology*, **52**, 1165–1172.

Borgan, Ø., Langholz, B., Samuelsen, S. O., Goldstein, L. and Pogoda, J. (2000). Exposure stratified case-cohort designs. *Lifetime Data Analysis*, **6**, 39–58.

Carstensen, B., Plummer, M., Laara, E., Hills, M. (2017). Epi: A package for statistical analysis in epidemiology. R package version 2.15. URL `https://CRAN.R-project.org/package=Epi`

Keogh, R. H. and Cox, D. R. (2014). *Case-control studies*. Cambridge University Press.

Kupper, L. L., McMichael, A. J. and Spirtas, R. (1975). A hybrid epidemiologic study design useful in estimating relative risk. *Journal of the American Statistical Association*, **70**, 524–528.

Liddell, F. D. K., McDonald, J. C., Thomas, D. C. and Cunliffe, S. V. (1977). Methods of cohort analysis: appraisal by application to asbestos mining. *Journal of the Royal Statistical Society* A, **140**, 469–491.

References

Onland-Moret, N. C., van der A, D. L., van der Schouw, Y. T., Buschers, W., Elias, S. G., van Gils, C. H. et al. (2007). Analysis of case-cohort data: a comparison of different methods. *Journal of Clinical Epidemiology*, **60**, 350–355.

Pai, J. K., Pischon, T., Ma, J., Manson, J. E., Hankinson, S. E., Joshipura, K., Curhan, G. C., Rifai, N., Cannuscio, C. C., Stampfer, M. J. and Rimm, E. B. (2004). Inflammatory markers and the risk of coronary heart disease in men and women. *New England Journal of Medicine*, **16**, 2599–2610.

Prentice, R. L. and Breslow, N. E. (1978). Retrospective studies and failure time models. *Biometrika*, **65**, 153–158.

Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, **73**, 1–11.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Self, S. G. and Prentice, R. L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *Annals of Statistics*, **16**, 64–81.

Therneau, T. (2015). survival: A package for survival analysis in S. Version 2.38. URL `http://CRAN.R-project.org/package=survival`