Computational Methods For Case-Cohort Studies

Sahir Rai Bhatnagar

Queen's University

November 27, 2012

Cohort studies

- All participants provide a wide range of information at time of recruitment e.g. detailed dietary questionnaires and blood and urine samples
- Because of large numbers and cost of analysing the biological specimens or genotyping, these resources are often not analysed in detail at the time but are stored for future use
- This design is expensive, inefficient for rare outcomes, long follow-up period needed, large sample size needed

Cohort studies

- All participants provide a wide range of information at time of recruitment e.g. detailed dietary questionnaires and blood and urine samples
- Because of large numbers and cost of analysing the biological specimens or genotyping, these resources are often not analysed in detail at the time but are stored for future use
- This design is expensive, inefficient for rare outcomes, long follow-up period needed, large sample size needed

Case-Cohort: A more efficient design

- A **random** sample of participants are selected from full cohort at baseline
- Detailed exposure information (covariates) can then be retrieved for

Case-Cohort: A more efficient design

- A **random** sample of participants are selected from full cohort at baseline
- Detailed exposure information (covariates) can then be retrieved for
  - this subcohort

Case-Cohort: A more efficient design

- A **random** sample of participants are selected from full cohort at baseline
- Detailed exposure information (covariates) can then be retrieved for
  - this subcohort
  - everyone in the full cohort who develop the disease of interest

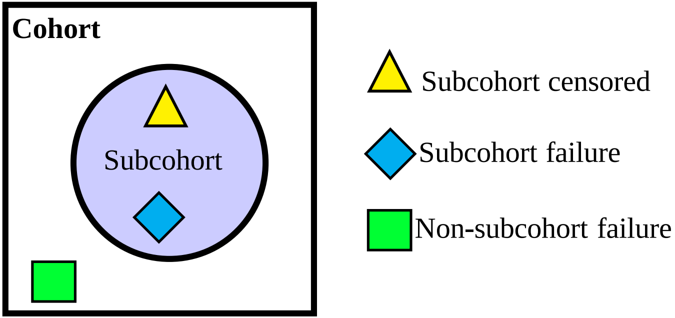
Case-Cohort: A more efficient design

- A **random** sample of participants are selected from full cohort at baseline
- Detailed exposure information (covariates) can then be retrieved for
  - this subcohort
  - everyone in the full cohort who develop the disease of interest
- Key feature: inclusion of all cases that occur in the cohort

Case-Cohort: A more efficient design

- A **random** sample of participants are selected from full cohort at baseline
- Detailed exposure information (covariates) can then be retrieved for
  - this subcohort
  - everyone in the full cohort who develop the disease of interest
- Key feature: inclusion of all cases that occur in the cohort

Case-Cohort Design



Objective

Purpose of this presentation

1. Explain and promote the case-cohort design
2. Show that it's not as difficult as the literature says to compute accurate estimates

Objective

Purpose of this presentation

1. Explain and promote the case-cohort design
2. Show that it's not as difficult as the literature says to compute accurate estimates

An Example

Description of the analysed dataset

- Simple and age at first exposure stratified case-cohort samples drawn from a cohort of 1741 female patients who were discharged from two tuberculosis sanatoria in Massachusetts between 1930 and 1956 to investigate **breast cancer risk** and **radiation exposure** due to fluoroscopy

An Example

Description of the analysed dataset

- Simple and age at first exposure stratified case-cohort samples drawn from a cohort of 1741 female patients who were discharged from two tuberculosis sanatoria in Massachusetts between 1930 and 1956 to investigate **breast cancer risk** and **radiation exposure** due to fluoroscopy
- Radiation doses were estimated for those women who received radiation exposure to the chest from the X-ray fluoroscopy lung examination

An Example

Description of the analysed dataset

- Simple and age at first exposure stratified case-cohort samples drawn from a cohort of 1741 female patients who were discharged from two tuberculosis sanatoria in Massachusetts between 1930 and 1956 to investigate **breast cancer risk** and **radiation exposure** due to fluoroscopy
- Radiation doses were estimated for those women who received radiation exposure to the chest from the X-ray fluoroscopy lung examination
- The remaining women received treatments that did not require fluoroscopic monitoring and were radiation unexposed

An Example

Description of the analysed dataset

- Simple and age at first exposure stratified case-cohort samples drawn from a cohort of 1741 female patients who were discharged from two tuberculosis sanatoria in Massachusetts between 1930 and 1956 to investigate **breast cancer risk** and **radiation exposure** due to fluoroscopy
- Radiation doses were estimated for those women who received radiation exposure to the chest from the X-ray fluoroscopy lung examination
- The remaining women received treatments that did not require fluoroscopic monitoring and were radiation unexposed
- 75 breast cancer cases were identified with 54 exposed and 21 unexposed

An Example

Description of the analysed dataset

- Simple and age at first exposure stratified case-cohort samples drawn from a cohort of 1741 female patients who were discharged from two tuberculosis sanatoria in Massachusetts between 1930 and 1956 to investigate **breast cancer risk** and **radiation exposure** due to fluoroscopy
- Radiation doses were estimated for those women who received radiation exposure to the chest from the X-ray fluoroscopy lung examination
- The remaining women received treatments that did not require fluoroscopic monitoring and were radiation unexposed
- 75 breast cancer cases were identified with 54 exposed and 21 unexposed
- 100 subjects were randomly sampled without replacement

An Example

Description of the analysed dataset

- Simple and age at first exposure stratified case-cohort samples drawn from a cohort of 1741 female patients who were discharged from two tuberculosis sanatoria in Massachusetts between 1930 and 1956 to investigate **breast cancer risk** and **radiation exposure** due to fluoroscopy
- Radiation doses were estimated for those women who received radiation exposure to the chest from the X-ray fluoroscopy lung examination
- The remaining women received treatments that did not require fluoroscopic monitoring and were radiation unexposed
- 75 breast cancer cases were identified with 54 exposed and 21 unexposed
- 100 subjects were randomly sampled without replacement

Advantages

- Exposure precedes outcome, while smaller scale reduces cost and effort
- In outbreak situations, multiple outcomes can be studied using only one sample of controls

Advantages

- Exposure precedes outcome, while smaller scale reduces cost and effort
- In outbreak situations, multiple outcomes can be studied using only one sample of controls

Advantages

- Exposure precedes outcome, while smaller scale reduces cost and effort
- In outbreak situations, multiple outcomes can be studied using only one sample of controls

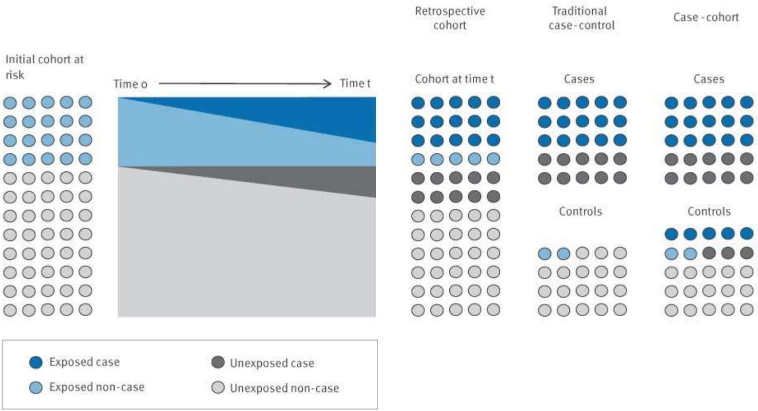
Challenges

- Theoretically computationally difficult to compute variance estimates
- Because of such biased sampling with regard to case-status, risk estimation using the ordinary partial likelihood is not appropriate

Challenges

- Theoretically computationally difficult to compute variance estimates
- Because of such biased sampling with regard to case-status, risk estimation using the ordinary partial likelihood is not appropriate

Comparing three study designs



Waroux et al.,2012

Cox Proportional Hazards Model

First lets consider a relative risk regression model *(Cox, 1972)*

Cox PH Model

$$\lambda\left\{t; Z(u), 0 \leq u \leq t\right\} = \lambda_0(t) r\left\{X(t)\beta\right\}$$

Cox Proportional Hazards Model

First lets consider a relative risk regression model *(Cox, 1972)*

Cox PH Model

$$\lambda\{t; Z(u), 0 \leq u \leq t\} = \lambda_0(t) r\{X(t)\beta\}$$

- $\lambda(t)$: failure rate of interest at time t for a subject

Cox Proportional Hazards Model

First lets consider a relative risk regression model *(Cox, 1972)*

Cox PH Model

$$\lambda\{t; Z(u), 0 \leq u \leq t\} = \lambda_0(t) r\{X(t)\beta\}$$

- $\lambda(t)$: failure rate of interest at time t for a subject
- $\{Z(u); 0 \leq u < t\}$:preceding covariate history

Cox Proportional Hazards Model

First lets consider a relative risk regression model *(Cox, 1972)*

Cox PH Model

$$\lambda\left\{t; Z(u), 0 \leq u \leq t\right\} = \lambda_0(t) r\left\{X(t)\beta\right\}$$

- $\lambda(t)$: failure rate of interest at time t for a subject
- $\{Z(u); 0 \leq u < t\}$:preceding covariate history
- $r(x)$: is a fixed function with $r(0) = 1$ e.g. $r(x) = exp\left\{x\right\}$

Cox Proportional Hazards Model

First lets consider a relative risk regression model *(Cox, 1972)*

Cox PH Model

$$\lambda\{t; Z(u), 0 \leq u \leq t\} = \lambda_0(t) r\{X(t)\beta\}$$

- $\lambda(t)$: failure rate of interest at time t for a subject
- $\{Z(u); 0 \leq u < t\}$:preceding covariate history
- $r(x)$: is a fixed function with $r(0) = 1$ e.g. $r(x) = exp\{x\}$
- $X(t)$: row p-vector consisting of functions of $Z(u)$

Cox Proportional Hazards Model

First lets consider a relative risk regression model *(Cox, 1972)*

Cox PH Model

$\lambda\left\{t; Z(u), 0 \leq u \leq t\right\} = \lambda_0(t) r\left\{X(t)\beta\right\}$

- $\lambda(t)$: failure rate of interest at time t for a subject
- $\{Z(u); 0 \leq u < t\}$:preceding covariate history
- $r(x)$: is a fixed function with $r(0) = 1$ e.g. $r(x) = exp\left\{x\right\}$
- $X(t)$: row p-vector consisting of functions of $Z(u)$
- β: column p-vector of regression parameters to be estimated

Cox Proportional Hazards Model

First lets consider a relative risk regression model *(Cox, 1972)*

Cox PH Model

$$\lambda\left\{t; Z(u), 0 \leq u \leq t\right\} = \lambda_0(t) r\left\{X(t)\beta\right\}$$

- $\lambda(t)$: failure rate of interest at time t for a subject
- $\{Z(u); 0 \leq u < t\}$:preceding covariate history
- $r(x)$: is a fixed function with $r(0) = 1$ e.g. $r(x) = exp\left\{x\right\}$
- $X(t)$: row p-vector consisting of functions of $Z(u)$
- β: column p-vector of regression parameters to be estimated
- $\lambda_0(t)$: baseline hazard function

Cox Proportional Hazards Model

First lets consider a relative risk regression model *(Cox, 1972)*

Cox PH Model

$$\lambda\left\{t; Z(u), 0 \leq u \leq t\right\} = \lambda_0(t) r\left\{X(t)\beta\right\}$$

- $\lambda(t)$: failure rate of interest at time t for a subject
- $\{Z(u); 0 \leq u < t\}$:preceding covariate history
- $r(x)$: is a fixed function with $r(0) = 1$ e.g. $r(x) = exp\left\{x\right\}$
- $X(t)$: row p-vector consisting of functions of $Z(u)$
- β: column p-vector of regression parameters to be estimated
- $\lambda_0(t)$: baseline hazard function

Exact and approximate pseudolikelihood estimators

- Indicator whether subject is at risk at time t

$$\tilde{\mathcal{L}}(\beta) = \prod_{i=1}^{n} \prod_{t} \left[\frac{\exp\{\beta Z_i(t)\}}{\sum_{k \in \tilde{\mathfrak{R}}_i(t)} \exp\{\beta Z_k(t)\}} \right]^{dN_i(t)} \tag{1}$$

- The contribution of a failure by subject i at time t

Exact and approximate pseudolikelihood estimators

- Indicator whether subject is at risk at time t

$$\tilde{\mathcal{L}}(\beta) = \prod_{i=1}^{n} \prod_{t} \left[\frac{\exp\{\beta Z_i(t)\}}{\sum_{k \in \tilde{\Re}_i(t)} \exp\{\beta Z_k(t)\}} \right]^{dN_i(t)} \tag{1}$$

- The contribution of a failure by subject i at time t
- Sum of all subcohort nonfailures at risk at time t including the failure by subject i

Exact and approximate pseudolikelihood estimators

- Indicator whether subject is at risk at time t

$$\tilde{\mathcal{L}}(\beta) = \prod_{i=1}^{n} \prod_{t} \left[\frac{\exp\{\beta Z_i(t)\}}{\sum_{k \in \tilde{\Re}_i(t)} \exp\{\beta Z_k(t)\}} \right]^{dN_i(t)} \quad (1)$$

- The contribution of a failure by subject i at time t
- Sum of all subcohort nonfailures at risk at time t including the failure by subject i
- Exact: $\tilde{\Re}_i(t) = (C \cup \{i\}) \cap \Re(t)$
- Approximate: $\tilde{\Re}_i(t) = C \cap \Re(t)$, where C is the subcohort

Exact and approximate pseudolikelihood estimators

- Indicator whether subject is at risk at time t

$$\tilde{\mathcal{L}}(\beta) = \prod_{i=1}^{n} \prod_{t} \left[\frac{\exp\{\beta Z_i(t)\}}{\sum_{k \in \tilde{\Re}_i(t)} \exp\{\beta Z_k(t)\}} \right]^{dN_i(t)} \tag{1}$$

- The contribution of a failure by subject i at time t
- Sum of all subcohort nonfailures at risk at time t including the failure by subject i
- Exact: $\tilde{\Re}_i(t) = (C \cup \{i\}) \cap \Re(t)$
- Approximate: $\tilde{\Re}_i(t) = C \cap \Re(t)$, where C is the subcohort

Exact and approximate pseudolikelihood estimators

- The unique sampling approach i.e. over selecting cases, leads to a **pseudolikelihood** rather than the usual partial likelihood
- Analysis must adjust for bias introduced in the distributions of covariates used in calculating the denominator of the pseudolikelihood

Exact and approximate pseudolikelihood estimators

- The unique sampling approach i.e. over selecting cases, leads to a **pseudolikelihood** rather than the usual partial likelihood
- Analysis must adjust for bias introduced in the distributions of covariates used in calculating the denominator of the pseudolikelihood
- Bias incurred by including cases outside the subcohort is corrected by not allowing those cases to contribute to risk sets other than their own

Exact and approximate pseudolikelihood estimators

- The unique sampling approach i.e. over selecting cases, leads to a **pseudolikelihood** rather than the usual partial likelihood
- Analysis must adjust for bias introduced in the distributions of covariates used in calculating the denominator of the pseudolikelihood
- Bias incurred by including cases outside the subcohort is corrected by not allowing those cases to contribute to risk sets other than their own
- We will focus our attention the **exact** approach rather than the approximate

Exact and approximate pseudolikelihood estimators

- The unique sampling approach i.e. over selecting cases, leads to a **pseudolikelihood** rather than the usual partial likelihood
- Analysis must adjust for bias introduced in the distributions of covariates used in calculating the denominator of the pseudolikelihood
- Bias incurred by including cases outside the subcohort is corrected by not allowing those cases to contribute to risk sets other than their own
- We will focus our attention the **exact** approach rather than the approximate

Approximate Variance of $\hat{\beta}$

Therneau and Li (1999) solved the variance estimation problem proposing the following approximation

$$\hat{I}^{-1} + \frac{m(n-m)}{n}\mathrm{Cov}\, D_C \tag{2}$$

- $\hat{I}^{-1}$: estimated covariance matrix of the parameter estimates (Inverse of Fisher Information matrix)

Approximate Variance of $\hat{\beta}$

Therneau and Li (1999) solved the variance estimation problem proposing the following approximation

$$\hat{I}^{-1} + \frac{m(n-m)}{n} \mathrm{Cov}\, D_C \qquad (2)$$

- $\hat{I}^{-1}$: estimated covariance matrix of the parameter estimates (Inverse of Fisher Information matrix)
- **n**: size of full cohort

Approximate Variance of $\hat{\beta}$

Therneau and Li (1999) solved the variance estimation problem proposing the following approximation

$$\hat{I}^{-1} + \frac{m(n-m)}{n} \mathrm{Cov}\, D_C \tag{2}$$

- $\hat{I}^{-1}$: estimated covariance matrix of the parameter estimates (Inverse of Fisher Information matrix)
- **n**: size of full cohort
- **m**: size of subcohort

Approximate Variance of $\hat{\beta}$

Therneau and Li (1999) solved the variance estimation problem proposing the following approximation

$$\hat{I}^{-1} + \frac{m(n-m)}{n} \mathrm{Cov}\, D_C \qquad (2)$$

- $\hat{I}^{-1}$: estimated covariance matrix of the parameter estimates (Inverse of Fisher Information matrix)
- **n**: size of full cohort
- **m**: size of subcohort
- $\mathrm{Cov} D_C$: empirical covariance matrix of *dfbeta* residuals from subcohort members

Approximate Variance of $\hat{\beta}$

Therneau and Li (1999) solved the variance estimation problem proposing the following approximation

$$\hat{I}^{-1} + \frac{m(n-m)}{n} \mathrm{Cov}\, D_C \tag{2}$$

- $\hat{I}^{-1}$: estimated covariance matrix of the parameter estimates (Inverse of Fisher Information matrix)
- **n**: size of full cohort
- **m**: size of subcohort
- $\mathbf{Cov} D_C$: empirical covariance matrix of *dfbeta* residuals from subcohort members

Dfbeta residuals

Are the approximate changes in the parameter estimates $(\hat{\beta} - \hat{\beta}_{(j)})$ when the j^{th} observation is omitted. These variables are a weighted transform of the score residual variables and are useful in assessing local influence and in computing approximate and robust variance estimates.

Procedure for creating analytic dataset

Steps

1. Each subcohort non-failure contributes one line of data to the analytic data set as censored observations

Procedure for creating analytic dataset

Steps

1. Each subcohort non-failure contributes one line of data to the analytic data set as censored observations
2. A non-subcohort failure contributes no information prior to the failure time so one line of data is contributed to the analytic data set as a failure but only at the failure time

Procedure for creating analytic dataset

Steps

1. Each subcohort non-failure contributes one line of data to the analytic data set as censored observations
2. A non-subcohort failure contributes no information prior to the failure time so one line of data is contributed to the analytic data set as a failure but only at the failure time
3. A subcohort failure contributes two lines to the analytic data set:

Procedure for creating analytic dataset

Steps

1. Each subcohort non-failure contributes one line of data to the analytic data set as censored observations
2. A non-subcohort failure contributes no information prior to the failure time so one line of data is contributed to the analytic data set as a failure but only at the failure time
3. A subcohort failure contributes two lines to the analytic data set:
   - one line as a censored observation prior to the failure time

Procedure for creating analytic dataset

Steps

1. Each subcohort non-failure contributes one line of data to the analytic data set as censored observations
2. A non-subcohort failure contributes no information prior to the failure time so one line of data is contributed to the analytic data set as a failure but only at the failure time
3. A subcohort failure contributes two lines to the analytic data set:
   - one line as a censored observation prior to the failure time
   - and one line as a failure at the failure time

Procedure for creating analytic dataset

Steps

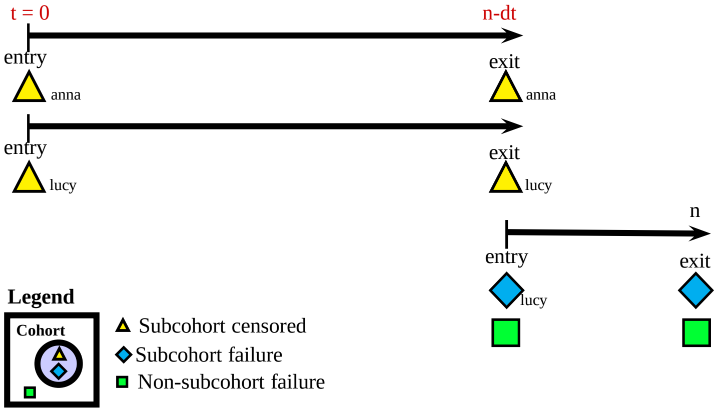
1. Each subcohort non-failure contributes one line of data to the analytic data set as censored observations
2. A non-subcohort failure contributes no information prior to the failure time so one line of data is contributed to the analytic data set as a failure but only at the failure time
3. A subcohort failure contributes two lines to the analytic data set:
   - one line as a censored observation prior to the failure time
   - and one line as a failure at the failure time
4. To create a time just before the exit time, an amount less than the precision of exit times given in the data is subtracted off from the actual failure time

Procedure for creating analytic dataset

Steps

1. Each subcohort non-failure contributes one line of data to the analytic data set as censored observations
2. A non-subcohort failure contributes no information prior to the failure time so one line of data is contributed to the analytic data set as a failure but only at the failure time
3. A subcohort failure contributes two lines to the analytic data set:
   - one line as a censored observation prior to the failure time
   - and one line as a failure at the failure time
4. To create a time just before the exit time, an amount less than the precision of exit times given in the data is subtracted off from the actual failure time

Graphic of how to create analytic dataset



Original Case Cohort dataset

Basic case-cohort data

Subject ID	Dose in rad	Age at exit (in years)	Age at entry (in years)	0-cens,1-subc fail 2-non-subc fail	1-249 rad	250+ rad	age at first exposure group [a]
2866	0.4525	71.269	34.0014	0	1	0	3
2787	0.00984	69.0294	31.7454	0	1	0	4
2702	0.05486	47.5948	36.5065	0	1	0	3
34	0	55.4387	14.9377	1	0	0	1
3064	0.12788	35.4825	25.6838	0	1	0	3
2766	1.62311	64.3559	30.5161	0	1	0	3
2344	1.0624	69.692	25.4127	0	1	0	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2698	0	42.3682	36.2026	2	0	0	4
2577	1.00338	50.9979	26.412	2	1	0	3
2348	1.30725	42.1246	24.1259	2	1	0	3
3106	0	55.2635	27.2635	2	0	0	3
2687	0	47.7563	23.2553	2	0	0	3
3018	1.6723	50.0014	38.8337	2	1	0	4

[a] $1 :< 15, 2 : 15-19, 3 : 20-29, 4 : 30+$

Comparison

Original vs. Analytic dataset

Subject ID	Dose in rad	Age at exit (in years)	Age at entry (in years)	0-cens,1-subc fail 2-non-subc fail	1-249 rad	250+ rad	age at first exposure group	an_entry	an_exit	an_ind
34	0	55.4387	14.9377	1	0	0	1			
2344	1.0624	69.692	25.4127	0	1	0	3			
2687	0	47.7563	23.2553	2	0	0	3			
34	0	55.4387	14.9377	1	0	0	1	14.9377	55.4386	0
34	0	55.4387	14.9377	1	0	0	1	55.4386	55.4387	1
2344	1.0624	69.692	25.4127	0	1	0	3	25.4127	69.6919	0
2687	0	47.7563	23.2553	2	0	0	3	47.7562	47.7563	1

Comparison

Original vs. Analytic dataset

Subject ID	Dose in rad	Age at exit (in years)	Age at entry (in years)	0-cens,1-subc fail 2-non-subc fail	1-249 rad	250+ rad	age at first exposure group	an_entry	an_exit	an_ind
34	0	55.4387	14.9377	1	0	0	1			
2344	1.0624	69.692	25.4127	0	1	0	3			
2687	0	47.7563	23.2553	2	0	0	3			
34	0	55.4387	14.9377	1	0	0	1	14.9377	55.4386	0
34	0	55.4387	14.9377	1	0	0	1	55.4386	55.4387	1
2344	1.0624	69.692	25.4127	0	1	0	3	25.4127	69.6919	0
2687	0	47.7563	23.2553	2	0	0	3	47.7562	47.7563	1

SAS Code

```
proc phreg data=analytic;
  model an_exit*an_ind(0)  = dcat1 dcat2 /
   entry=an_entry covb;
  output out=dfbetas dfbeta= dfb_dcat1 dfb_dcat2;
  id id;
run;

proc corr data=dfbetas cov;
  var dfb_dcat1 dfb_dcat2;
  where an_ind eq 0;
run;
```

- covb: outputs the inverse information matrix $\hat{I}^{-1}$
- cov: outputs the covariance matrix of *dfbeta* residuals from subcohort members (WHERE an_ind $= 0$)

SAS Code

```
proc phreg data=analytic;
  model an_exit*an_ind(0)  = dcat1 dcat2 /
   entry=an_entry covb;
  output out=dfbetas dfbeta= dfb_dcat1 dfb_dcat2;
  id id;
run;

proc corr data=dfbetas cov;
  var dfb_dcat1 dfb_dcat2;
  where an_ind eq 0;
run;
```

- covb: outputs the inverse information matrix $\hat{I}^{-1}$
- cov: outputs the covariance matrix of *dfbeta* residuals from subcohort members (WHERE an_ind $= 0$)

Exact pseudolikelihood and Asymptotic variance

Analysis of Maximum Likelihood Estimates							
Parameter	**DF**	Parameter Estimate	Standard Error	χ^2	**Pr**$> \chi^2$	Hazard Ratio	**Label**
dcat1	1	0.6572	0.26117	6.332	0.0119	1.929	1-249 rad
dcat2	1	1.55325	0.50118	9.6051	0.0019	4.727	250+ rad

Estimated Covariance Matrix $(\times 10^{-2})$			
Parameter		**dcat1**	**dcat2**
dcat1	**1-249 rad**	6.821	4.743
dcat2	**250+ rad**	4.743	25.118

Estimated Covariance Matrix of the dfbeta residuals $(\times 10^{-4})$			
Parameter		**dfb_dcat1**	**dfb_dcat2**
dfb_dcat1	**difference in the parameter for dcat1**	5.487	2.998
dfb_dcat2	**difference in the parameter for dcat2**	2.998	47.878

Time-dependent covariates

- The partial likelihood of Cox also allows time-dependent explanatory variables
- An explanatory variable is time-dependent if its value for any given individual can change over time

Time-dependent covariates

- The partial likelihood of Cox also allows time-dependent explanatory variables
- An explanatory variable is time-dependent if its value for any given individual can change over time
- We introduce a latency variable **lat15** indicating 15 years since last fluoroscopy

Time-dependent covariates

- The partial likelihood of Cox also allows time-dependent explanatory variables
- An explanatory variable is time-dependent if its value for any given individual can change over time
- We introduce a latency variable **lat15** indicating 15 years since last fluoroscopy

Difficulties in programming

- Most software can account for time-dependent covariates for rate ratio estimation, however none can compute *dfbeta* residuals for these time-dependent covariates
- Thus it is not possible to compute the robust or asymptotic variance estimators for case-cohort data

Difficulties in programming

- Most software can account for time-dependent covariates for rate ratio estimation, however none can compute *dfbeta* residuals for these time-dependent covariates
- Thus it is not possible to compute the robust or asymptotic variance estimators for case-cohort data

Proposed solution

- Software can be "tricked" to accommodate time-dependent covariates by organizing the case-cohort data into risk sets
- Has the structure of individually matched case-control data with a risk set formed at each failure time

Proposed solution

- Software can be "tricked" to accommodate time-dependent covariates by organizing the case-cohort data into risk sets
- Has the structure of individually matched case-control data with a risk set formed at each failure time
- **Case**: is the failure at a specific failure time
- **Controls**: are all those still at risk at the case failure time

Proposed solution

- Software can be “tricked” to accommodate time-dependent covariates by organizing the case-cohort data into risk sets
- Has the structure of individually matched case-control data with a risk set formed at each failure time
- **Case**: is the failure at a specific failure time
- **Controls**: are all those still at risk at the case failure time

Analytic dataset

Analytic Dataset for time dependent covariates

Caseid	set_no	rstime	rsentry	Subject ID	Dose in rad	Age at exit (in years)	Age at entry (in years)	0-cens,1-subc fail 2-non-subc fail	1-249 rad	250+ rad	age at first exposure group	ccohentry	cc	latency	lat15
22	1	25.4292	25.4291	22	0.61714	25.4292	17.1773	1	1	0	1	17.1773	1	8.2519	0
22	1	25.4292	25.4291	2958	4.13045	33.6016	15.8303	0	0	1	1	15.8303	0	9.5989	0
22	1	25.4292	25.4291	295	0.58148	51.833	17.5496	0	1	0	2	17.5496	0	7.8795	0
22	1	25.4292	25.4291	261	0	52.8569	3.4771	0	0	0	1	3.4771	0	21.9521	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
22	1	25.4292	25.4291	34	0	55.4387	14.9377	1	0	0	1	14.9377	0	10.4914	0
22	1	25.4292	25.4291	334	1.15677	56.6543	18.3381	0	1	0	2	18.3381	0	7.091	0
22	1	25.4292	25.4291	2057	0	73.2402	20.8049	0	0	0	2	20.8049	0	4.6242	0
2350	33	47.7235	47.7234	2350	0.94436	47.7235	20.2218	2	1	0	2	47.7234	1	27.5017	1
2350	33	47.7235	47.7234	3043	0.92604	48.909	17.5414	0	1	0	1	17.5414	0	30.1821	1
2350	33	47.7235	47.7234	242	0.67137	49.1608	15.8795	0	1	0	1	15.8795	0	31.8439	1
2350	33	47.7235	47.7234	2244	0.00959	49.1828	16.9035	0	1	0	2	16.9035	0	30.82	1
2350	33	47.7235	47.7234	3150	0	50.4723	17.191	0	0	0	2	17.191	0	30.5325	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2350	33	47.7235	47.7234	3317	1.28865	78.4559	27.7755	0	1	0	3	27.7755	0	19.948	1
2350	33	47.7235	47.7234	3182	0.95419	81.0951	35.05	0	1	0	4	35.05	0	12.6735	0
2350	33	47.7235	47.7234	3258	0.82631	86.642	38.36	0	1	0	4	38.36	0	9.3634	0
2350	33	47.7235	47.7234	3198	0	87.1157	42.5435	0	0	0	4	42.5435	0	5.18	0
3085	75	77.86	77.86	3085	0	77.8645	43.0335	2	0	0	4	77.8644	1	34.83	1
3085	75	77.86	77.86	3317	1.28865	78.4559	27.7755	0	1	0	3	27.7755	0	50.09	1
3085	75	77.86	77.86	3182	0.95419	81.0951	35.05	0	1	0	4	35.05	0	42.81	1
3085	75	77.86	77.86	3258	0.82631	86.642	38.36	0	1	0	4	38.36	0	39.5	1
3085	75	77.86	77.86	3198	0	87.1157	42.5435	0	0	0	4	42.5435	0	35.32	1
3085	75	77.86	77.86	2477	0	89.9849	53.9001	0	0	0	4	53.9001	0	23.96	1

SAS Code

```
proc phreg data=pclib.td_analytic nosummary;
  model rstime*cc(0) = dcat1 dcat2 lat15
    / entry=rsentry covb;
  output out=dfbetas dfbeta= dfb_dcat1 dfb_dcat2 dfb_lat15;
  id id;
run;

proc summary data=dfbetas sum;
   class id;
   var dfb_dcat1 dfb_dcat2 dfb_lat15;
   output out=summed sum=dfb_dcat1 dfb_dcat2 dfb_lat15;
   where cc eq 0;
run;

proc corr data=summed cov;
  var dfb_dcat1 dfb_dcat2 dfb_lat15;
run;
```

Exact pseudolikelihood estimators

Analysis of Maximum Likelihood Estimates

Parameter	**DF**	Parameter Estimate	Standard Error	χ^2	**Pr**$>\chi^2$	Hazard Ratio	**Label**
dcat1	1	0.65709	0.26112	6.3325	0.0119	1.929	1-249 rad
dcat2	1	1.68786	0.50750	11.0610	0.0009	4.727	250+ rad
lat15	1	0.61486	0.36062	2.9071	0.0882	1.849	

Stratification by age at first exposure

- It is quite possible that age is confounding the main effects of the covariates
- To control for confounding we stratify by age at first exposure group
- Each stratum (s) contributes independently to the pseudolikelihood
- the asymptotic variance is given by

$$\hat{I}^{-1} + \sum_{s} \frac{m_s(n_s - m_s)}{n_s} \mathrm{Cov}\, D_{C_s} \tag{3}$$

Stratification

Age Stratified Groups	
Age	**Group number**
<15	1
15-19	2
20-29	3
30+	4

SAS Code

```
proc phreg data=analytic;
  model an_exit*an_ind(0) = dcat1 dcat2 / entry=an_entry covb;
  output out=dfbetas dfbeta= dfb_dcat1 dfb_dcat2;
  strata agefirstgr;
  id id;
run;

proc corr data=dfbetas cov;
  var dfb_dcat1 dfb_dcat2;
  by agefirstgr;
  where an_ind eq 0;
run;
```

Exact pseudolikelihood estimators

Stratified

Analysis of Maximum Likelihood Estimates							
Parameter	**DF**	Parameter Estimate	Standard Error	χ^2	**Pr**$> \chi^2$	Hazard Ratio	**Label**
dcat1	1	0.5938	0.27148	4.7838	0.0287	1.811	1-249 rad
dcat2	1	0.9349	0.51737	3.2655	0.0708	2.547	250+ rad

Unstratified

Analysis of Maximum Likelihood Estimates							
Parameter	**DF**	Parameter Estimate	Standard Error	χ^2	**Pr**$> \chi^2$	Hazard Ratio	**Label**
dcat1	1	0.6572	0.26117	6.332	0.0119	1.929	1-249 rad
dcat2	1	1.55325	0.50118	9.6051	0.0019	4.727	250+ rad

Summary

1. Efficiency and benefits of case-cohort design
2. Take advantage of available software

Summary

1. Efficiency and benefits of case-cohort design
2. Take advantage of available software

References I

Bryan Langholz and Jenny Jiao
Computational methods for case cohort studies
Computational Statistics & Data Analysis, 51:2007

J. Cologne et al.
Conventional case-cohort design and analysis for studies of interaction
International Journal of Epidemiology, 41:2012

Therneau, T and Li, H
Computing the Cox model for case cohort designs
Lifetime data analysis, 2(5):1999

References II

Penny Webb and Chris Bain.
Essential Epidemiology.
Cambridge University Press, 2nd edition, 2011.

O Le Polain de Waroux
The case-cohort design in outbreak investigations
Euro surveillance, 17(25):2012