Accounting For A Non-Ignorable Tracing Mechanism In A Retrospective Breast Cancer Cohort Study

Andrew Titman, Gillian Lancaster
Lancaster University
Katie Carmichael and Diane Scutt
University of Liverpool.

Overview

- Breast cancer study
- Tracing problem
- Possible methods
  - Exclude some patients
  - Pseudo-likelihood method
- Results
- Conclusions

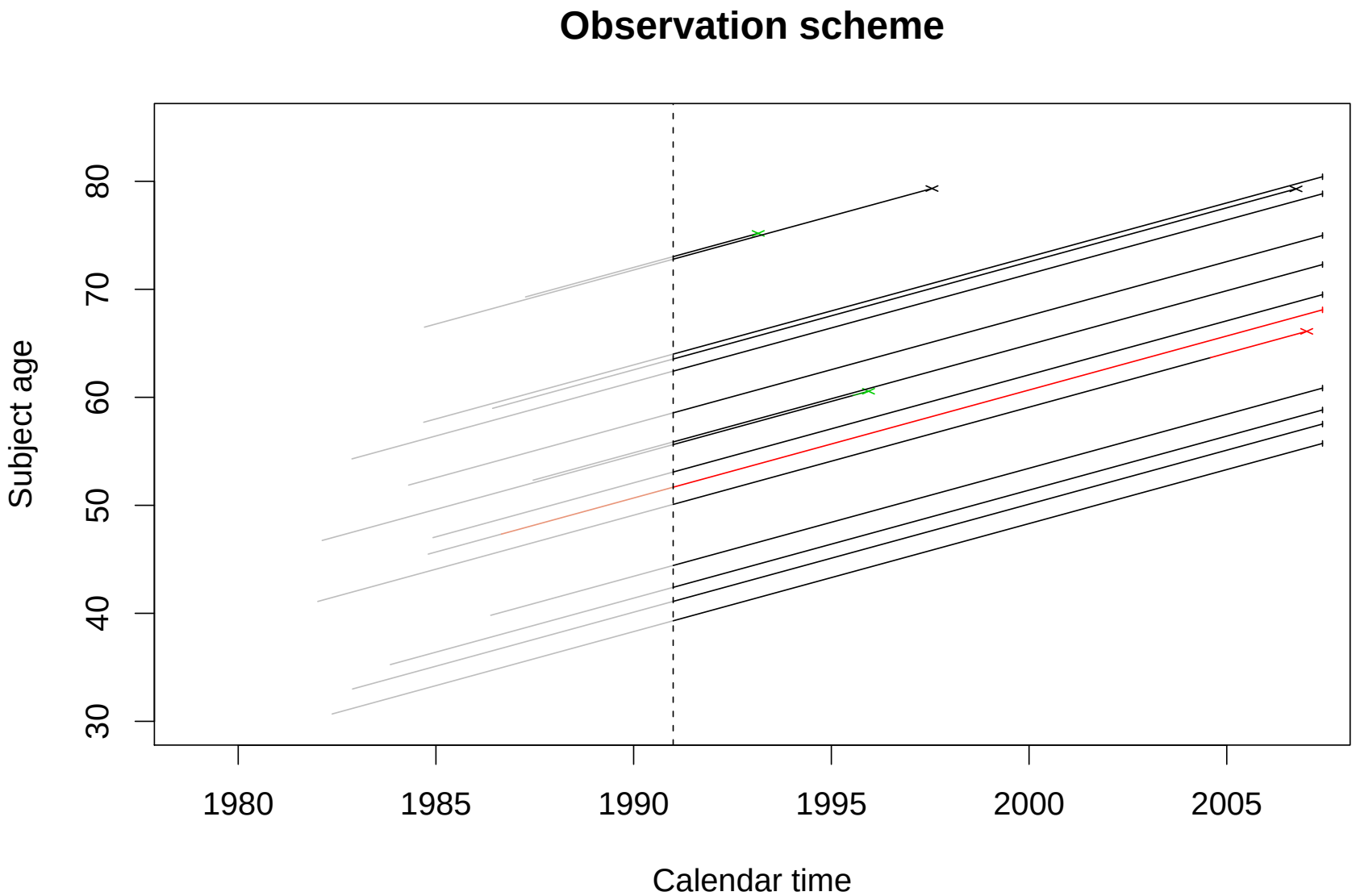
Breast cancer study

- Around 15000 self-selected women in the Merseyside area answered a detailed questionnaire and undertook a mammogram between 1979 and 1988.
- Ages ranged from 21-79, though vast majority aged 30-70.
- Questionnaire covered many potential risk factors, e.g.
  - Family history of breast cancer
  - Age at menarche
  - Previous breast biopsies
  - Height and weight.
- Mammogram allows:
  - Diagnosis of parenchymal pattern types
  - Assessment of breast volume asymmetry
- Objective to assess factors affecting onset rate of breast cancer.

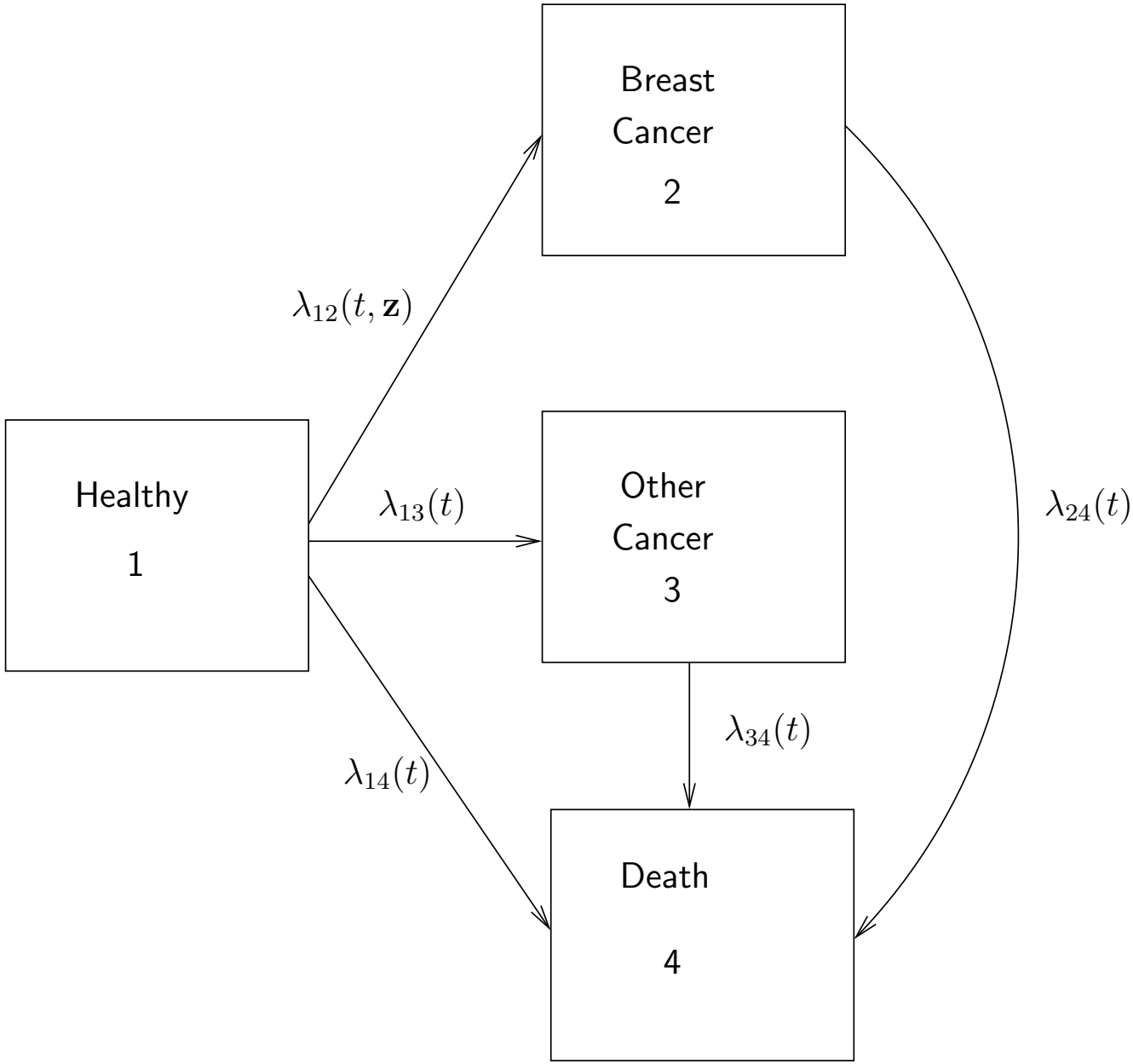
Retrospective tracing

- In 2007 women from the questionnaire were traced.
  - Information on all cancers.
  - Mortality
  - Emigration
- Tracing of women incomplete
  - Women only traced if on central health database (set up in 1991).
  - Women who died before 1991 would not be traced.

Observation scheme



Overall process



General problem

- Data from a multi-state model.
- Interested in estimating $\lambda_{12}(t, \mathbf{z})$ - rate of onset of a particular illness from healthy.
  - e.g. assume $\lambda_{12}(t, \mathbf{z}) = \lambda_{12}(t) \exp(\beta^T \mathbf{z})$
- Suppose there is a population of individuals, $i = 1, \ldots, N$.
- Subject i is included in the dataset provided they survive to a *truncation time* t_u.
- Non-standard type of left truncation.

Purged Processes

Before time t_u the traced subjects follow a conditional or *purged* process (Hoem, 1969).

$$\tilde{\lambda}_{12}(t, \mathbf{z}) = \lambda_{12}(t, \mathbf{z})(1 - p_{24}(t, t_u; \mathbf{z}))/(1 - p_{14}(t, t_u; \mathbf{z}))$$

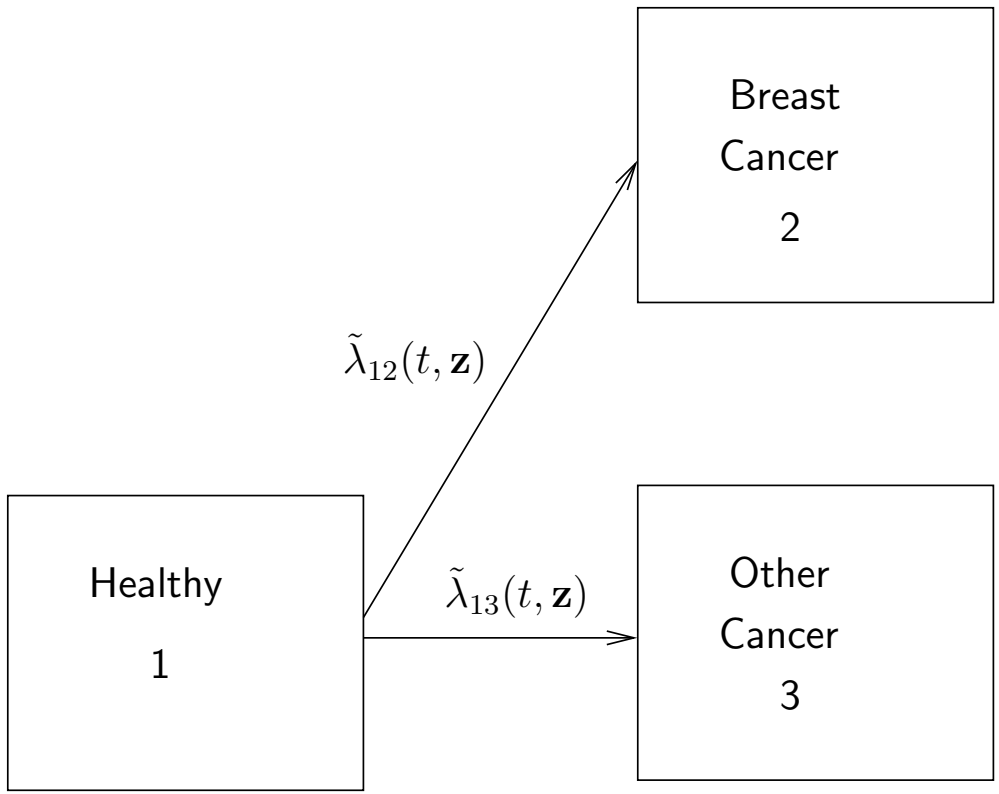
where $p_{r4}(t, t_u)$ represents probability of death by time t_u given in state r at time t. Distortion factor:

$$\rho(t) = (1 - p_{24}(t, t_u; \mathbf{z}))/(1 - p_{14}(t, t_u; \mathbf{z}))$$

- Unless illness has no effect on mortality, observed intensities will be distorted.
- Form of length bias but dependent on unknown parameters of interest e.g. λ_{12} and unknown nuisance parameters λ_{24}, λ_{34}.

Purged process

Observed process before time t_u:



Method 1: Post-truncation time analysis

- Exclude all women who had an event before 1991
- Results in exclusion of 268 patients
  - 72 breast cancer cases (out of 341)
  - 196 other cancer cases (out of 1023)
- Analysis is then a straightforward (left truncated) competing risks analysis
  - Separate Cox-proportional hazards models on cause-specific hazards.
- No need to model post-cancer survival.
- Loss of information.

Method 2: Pseudo-likelihood approach

- Kalbfleisch and Lawless (1988, Stats in Medicine), Copas and Farewell (2001, Biostatistics).
- Involves weighting log-likelihood contributions of traced individuals by estimated probabilities of being traced.
  - Idea is to produce a function with same expectation as the complete data likelihood.
- Assume a distribution, F_X, for the time (date) of entry into the study.
  - Assumptions about independence of covariates (including age at entry into study) and date of entry into study.
- Condition on the time between entry into study and first event.
- Estimate probability of tracing, treating time of entry into study as a random variable.

Pseudo-likelihood approach

$$pl_i(\theta) = \frac{l_i(\theta|T,\delta,Z)\Delta_i(X,T,\delta,Z)}{p_i(T,\delta,Z)}$$

where X time of entry into study, $X+T$ time to first event, δ indicator of event type, Z covariate values (including age at entry). Δ_i indicator of whether patient traced.

$$\begin{aligned}
\mathbb{E}(pl_i(\theta)) &= \mathbb{E}(\mathbb{E}(pl_i(\theta)|T,\delta,Z)) \\
&= \mathbb{E}\left(\int_{\mathcal{X}} \Delta_i(X,T,\delta,Z) dF_x(x) \frac{l_i(\theta)}{p_i(T,\delta,Z)}\right) \\
&= \mathbb{E}\left(p_i(T,\delta,Z)\frac{l_i(\theta)}{p_i(T,\delta,Z)}\right) \\
&= \mathbb{E}(l_i(T,\delta,Z))
\end{aligned}$$

Generalisation for partial-likelihood possible.

Calculation of tracing probabilities

- Censored or died from healthy at time T

$$p_i(T, Z) = \int_{\mathcal{X}} \mathbf{1}\{x + T > t_u\} dF_x(x)$$

  i.e. traced provided event occurred after t_u (e.g. 1991)
- Breast-cancer at time T

$$p_i(T, Z) = \int_{\mathcal{X}} \left[\mathbf{1}\{x + T > t_u\} + \mathbf{1}\{x + T \leq t_u\} S_{24}(t_u - x - T|\mathbf{z})\right] dF_x(x)$$

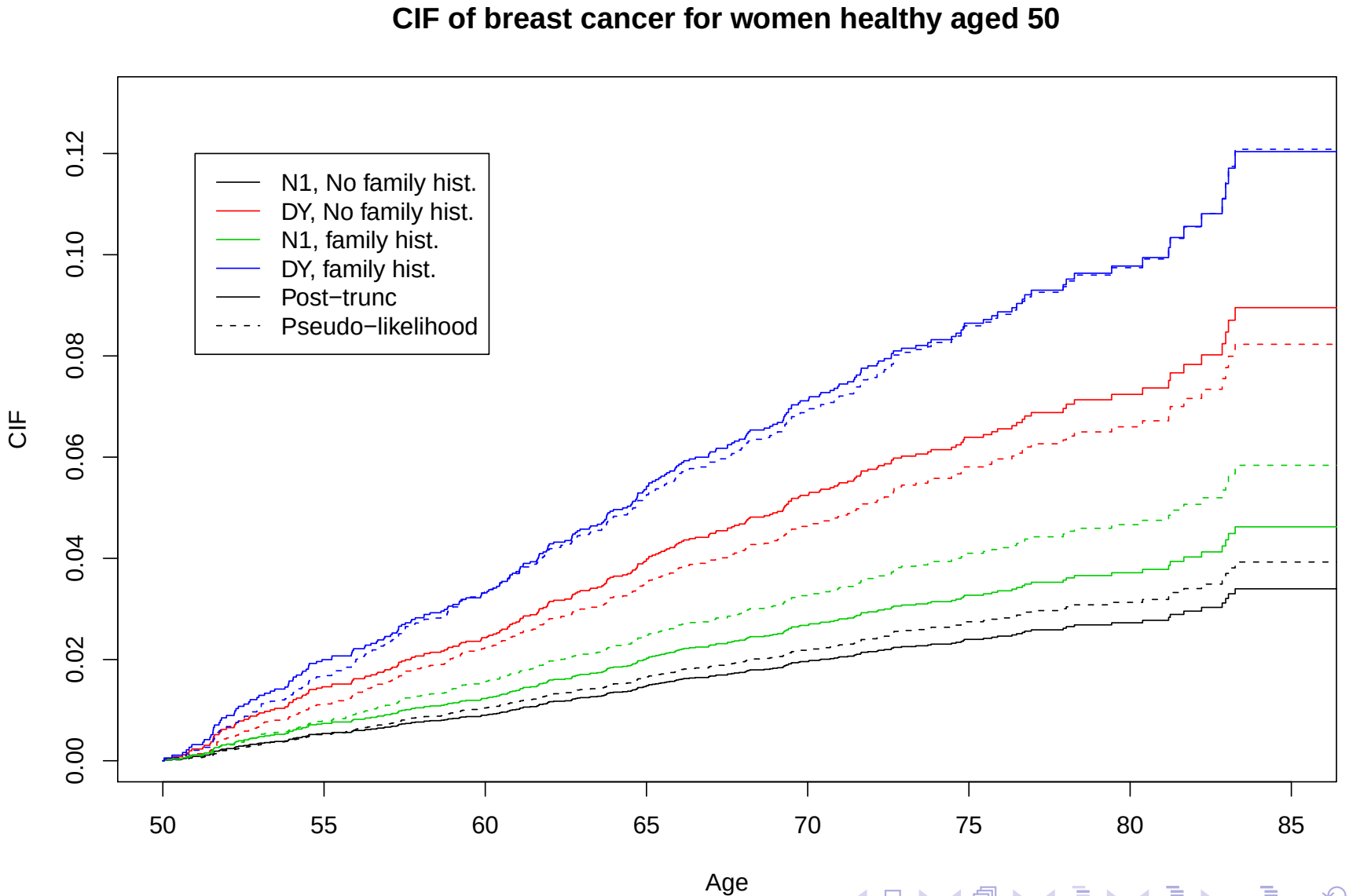
  i.e. traced provided event occurred after t_u or occurred before t_u but patient survived until t_u.
- S_{24} is survivor function from breast cancer which can be estimated independently.
- $F_x(x)$ distribution of study entry times which needs to be assumed or obtained from external data.
- Given tracing probabilities p_i, fitting is same as for Cox-PH models for survey data on biased samples (Binder, 1992).

Results

	Post-1991			Pseudo-likelihood		
	Est	SE(log)	p-value	Est	SE(log)	p-value
Rel. volume asym.	0.97	0.08	0.69	0.97	0.07	0.66
Partype: N1	1			1		
P1	1.95	0.27	0.01	1.49	0.21	0.06
P2	2.67	0.28	< 0.001	2.24	0.21	< 0.001
DY	2.72	0.26	< 0.001	2.22	0.20	< 0.001
Age at menarche	0.98	0.04	0.59	0.93	0.03	0.03
Height (per 10cm)	1.15	0.11	0.19	1.11	0.09	0.22
Family history	1.37	0.14	0.03	1.51	0.11	< 0.001
Biopsy	1.32	0.19	0.16	1.41	0.16	0.03

- Approximately a 30% reduction in standard errors.

Estimates of breast cancer cumulative incidence function



Conclusions: Methodological

- General agreement of point estimates between the methods
- Improved precision of estimates by using pre-1991 data.
- Pseudo-likelihood approach:
  - Only need to calculate survival functions to get weights
  - Weights not dependent on β.
  - Requires modelling of initiation times and post-illness survival.
  - Extra modelling may not be justified if amount of pre-truncation time data is small.

Conclusions: Clinical

- Breast volume asymmetry not associated with increased risk of breast cancer.
- Parenchymal patterns are a significant predictor.

References

- Hoem J.M. Purged and partial Markov chains. *Skandinavisk Aktvarietidsskrift* 1969; **52**: 147-155.
- Kalbfleisch J.D, Lawless J.F. Likelihood analysis of multi-state models for disease incidence and mortality. *Statistics in Medicine* 1988; **7**: 149-160.
- Copas A.J, Farewell V.T. Incorporating retrospective data into an analysis of time to illness. *Biostatistics* 2001; **2**: 1-12.
- Binder D.A. Fitting Cox's proportional hazards models from survey data. *Biometrika* 1992; **79**:139-147.

Acknowledgements:
Funded by CRUK Small Grant C24779/A9646

Calculation of baseline hazards

- Standard Breslow estimate

$$\hat{\Lambda}(t) = \sum_{i=1}^{N} \int_0^T \frac{Y_i(u)}{S^{(0)}(\beta, u)} dN_i(u)$$

where $S^{(0)}(\beta, u) = \sum_{j=1}^{N} Y_j(t) \exp\left(\beta^T \mathbf{z_j}(t)\right)$

- Pseudo-likelihood Breslow estimate

$$\hat{\Lambda}(t) = \sum_{i=1}^{N} p_i^{-1} \int_0^T \frac{Y_i(u)}{\tilde{S}^{(0)}(\beta, u)} dN_i(u)$$

where $\tilde{S}^{(0)}(\beta, u) = \sum_{j=1}^{N} p_j^{-1} Y_j(t) \exp\left(\beta^T \mathbf{z_j}(t)\right)$

Full likelihood

- When including pre-1991 data, need to account for tracing mechanism in the likelihood

$$l(\theta) = l^*(\theta) - \sum \log\left(1 - p_{1R}(t_{li}, t_{ui}; \theta)\right)$$

  where $l^*(\theta)$ is the likelihood under the assumption of an ignorable tracing mechanism and t_{li}, t_{ui} entry time and truncated time (i.e. age in 1991) for patient i.
- Full likelihood difficult to work with
  - General maximisation problems (c.f. standard Cox regression) unless assume parametric baseline intensities.
  - $p_{14}(t_l, t_u)$ hard to compute, even for parametric intensities unless Markov property assumed.
- Markov model with moderate number of piecewise constant intensities possible.