Accounting for informative tracing in a breast cancer cohort study

Andrew Titman

February 12, 2009
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
  - Family history of breast cancer
  - Age at menarche
  - Age at menopause
  - Previous breast biopsies
  - Number of pregnancies
  - Height and weight.
- Mammogram allows:
  - Diagnosis of parenchymal pattern types
  - Assessment of breast volume asymmetry
Methods of tracing

- Current study attempts to trace the women to determine effect of potential risk factors for breast cancer.
- Central Health Register Inquiry System (CHRIS) used to trace women
  - Only uses computer based records
  - Full manual trace too costly
  - Gives details of all cancers before and after 1991 for traced women.
- Tracing success rate rather poor (around 50%).
Methods of tracing

- Tracing success low for two main reasons
  - Women could not be positively matched in many cases because, for instance, only a woman’s first initial and not full name was recorded.
  - CHRIS system only became active in 1991, meaning that generally only women alive in 1991 were traced.

- First source of tracing error is likely to be ignorable.
- Second source potentially cause systematic bias.
Tracing bias

- Since only women alive in 1991 are traced and assuming cancer increases hazard of death
  - Would expect the observed cancer rate to be lower than the true rate in the period before 1991.
  - i.e. women who developed cancer before 1991 less likely to be traced.
Observation scheme
Choice of time scale

- Data has several inherent time scales
  - Calendar time
  - Time since recruitment in the study
  - Age of woman
- Since age has strong influence of hazard of cancer and death, this seem natural choice for timescale
  - Age effects accounted for non-parametrically.
Approaches to the tracing bias

▶ Post 1991 analysis
  ▶ Restrict analysis to women still healthy in 1991.
  ▶ Standard (left truncated) competing risks analysis then valid.
  ▶ But information loss (i.e. cancer cases pre-1991)
▶ Include pre-1991 data.
  ▶ Requires modelling of times to death from cancer.
  ▶ Multi-state modelling framework required.
  ▶ Needs additional assumptions
Post-1991 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
  - Women only enter the study conditional on being disease free in 1991.
- No need to model post-cancer survival.
Results on post-1991 analysis

Cox regression on the healthy to breast cancer cause-specific hazard:

<table>
<thead>
<tr>
<th>Rel. volume asym.</th>
<th>Est</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partype: N1</td>
<td>0</td>
<td>(-1.90, 0.63)</td>
</tr>
<tr>
<td>P1</td>
<td>0.72</td>
<td>(-0.03, 0.83)</td>
</tr>
<tr>
<td>P2</td>
<td>1.05</td>
<td>(0.18, 1.26)</td>
</tr>
<tr>
<td>DY</td>
<td>1.07</td>
<td>(0.54, 1.59)</td>
</tr>
<tr>
<td>Family history</td>
<td>0.31</td>
<td>(0.04, 0.59)</td>
</tr>
<tr>
<td>Biopsy</td>
<td>0.31</td>
<td>(-0.08, 0.69)</td>
</tr>
<tr>
<td>BMI</td>
<td>0.05</td>
<td>(0.00, 0.09)</td>
</tr>
<tr>
<td>Age menarche</td>
<td>-0.00</td>
<td>(-0.08, 0.07)</td>
</tr>
<tr>
<td>No. preg</td>
<td>-0.02</td>
<td>(-0.10, 0.06)</td>
</tr>
<tr>
<td>Age at menopause</td>
<td>0.03</td>
<td>(-0.03, 0.09)</td>
</tr>
</tbody>
</table>
Pre-and-post 1991 analysis

- Assume no cohort effect
  - Woman aged 35 in 1985 has same transition intensities as woman aged 35 in 1995.
- Explicit modelling of post-cancer survival rates needed
  - Depends on time since diagnosis
  - Depends on patient age
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.
  - \( p_{1R}(t_l, t_u) \) hard to compute, even for parametric intensities unless Markov property assumed.

- Markov model with moderate number of piecewise constant intensities possible.
Method 1: Pseudo-likelihood approach

- 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 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
Method 1: 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, \( \delta \) indicator of event type, \( Z \) covariate values (including age at entry). \( \Delta_i \) indicator of whether patient traced.

\[
\mathbb{E}(pl_i(\theta)) = \mathbb{E}(\mathbb{E}(pl_i(\theta) | T, \delta, Z)) \\
= \mathbb{E}\left( \int_{\mathcal{X}} \Delta_i(X, T, \delta, Z) dF_X(x) \frac{l_i(\theta)}{p_i(T, \delta, Z)} \right) \\
= \mathbb{E}\left( p_i(T, \delta, Z) \frac{l_i(\theta)}{p_i(T, \delta, Z)} \right) \\
= \mathbb{E}(l_i(T, \delta, Z))
\]
Method 1: Pseudo-likelihood approach

- Censored or died from healthy at time $T$
  \[ p_i(T, Z) = \int_{\mathcal{X}} \mathbf{1}\{x + T > t_c\} dF_x(x) \]
  i.e. traced provided event occurred after $t_c$ (e.g. 1991)

- Breast-cancer at time $T$
  \[ p_i(T, Z) = \int_{\mathcal{X}} (\mathbf{1}\{x + T > t_c\} + \mathbf{1}\{x + T \leq t_c\} S_{24}(t_c - x - T | z)) dF_x(x) \]
  i.e. traced provided event occurred after $t_c$ or occurred before $t_c$ but patient survived until $t_c$.

- $S_{24}$ 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.
Method 1: Estimation of study entry times

- For the breast cancer cohort study the questionnaire dates of all patients (whether traced or not) are available.
- It is therefore possible to estimate the distribution of study entry times.
- Recruitment appears to be close to uniformly distributed between around September 1979 and December 1987.
  - Assume $X \sim U(1979.75, 1988)$
Method 2: More direct estimation approach

- Tracing bias results in observed data pre-1991 coming from a conditional process.
- Hoem (1969) termed these purged processes.
- We can quantify the degree by which the tracing mechanism affects the observed transition intensity

\[ \tilde{\lambda}_{12}(t) = \frac{\lambda_{12}(t)(1 - p_{2R}(t, t_u))}{1 - p_{1R}(t, t_u)} \]

i.e. a distortion factor of
\[ \rho(t, t_u) = \frac{(1 - p_{2R}(t, t_u))}{(1 - p_{1R}(t, t_u))} \]
- \( \rho \) = ratio of survival probabilities to truncation time \( t_u \) if an event does or doesn’t occur at time \( t \).
- We expect \( \rho < 1 \) because expect event (cancer) to increase hazard of death.
Method 2: Example distortion factor
Method 2: Distortion factor approach

If the distortion factors were known and fixed then the partial likelihood could be written in the form

\[
I(\beta) = \sum_{i} \frac{\rho_{ii} \exp(\beta z_i)}{\sum_{r \in R_i} \rho_{ri} \exp(\beta z_r)}
\]

where \( \rho_{ri} = \rho(t_i, t_{ur}; z_r, \beta) \)

However, \( \rho \) actually depend on unknown \( \beta \) as well as baseline intensities.
Method 2: Iterative algorithm for estimation

1. Estimate $\beta$, baseline hazards and hence distortion factors using post-1991 analysis.

2. Use distortion factors as fixed hazard weights in a Cox-proportional hazards model to get a new estimate of $\beta$.

3. Update estimates of distortion factors based on new estimates of $\beta$ and the baseline hazards.

4. Return to step 2.

- In step 2. on the first iteration we are using unbiased estimates of the distortion factors.
- Effectiveness of the algorithm may depend on degree of dependence of distortion factors on $\beta$. 
## Comparison of results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est</td>
<td>95% CI</td>
<td>Est</td>
</tr>
<tr>
<td>P1</td>
<td>0</td>
<td></td>
<td>0.42</td>
</tr>
<tr>
<td>P2</td>
<td>0.40</td>
<td>(-0.03, 0.83)</td>
<td>0.84</td>
</tr>
<tr>
<td>DY</td>
<td>0.81</td>
<td>(0.41, 1.19)</td>
<td>0.83</td>
</tr>
<tr>
<td>Family history</td>
<td>0.40</td>
<td>(0.16, 0.63)</td>
<td>0.40</td>
</tr>
<tr>
<td>Biopsy</td>
<td>0.26</td>
<td>(-0.07, 0.60)</td>
<td>0.25</td>
</tr>
</tbody>
</table>
Estimates of breast cancer cumulative incidence function
Conclusions

- 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 $\beta$.
  - Requires modelling of initiation times.
- Distortion factor approach:
  - Requires calculation of $p_{1R}(t, t_u)$.
  - Iterative algorithm necessary for estimation.
  - Does not require modelling of initiation times.
Further work

- Sensitivity to different assumptions about post-cancer survival.
- Sensitivity of pseudo-likelihood approach to misspecified entry-time distribution.
- Comparison of efficiency between two methods.


Calculation of baseline hazards

- **Standard Breslow estimate**

\[
\hat{\Lambda}(t) = \sum_{i : t_i \leq t} \left( \sum_{r \in R_i} \exp (\beta^T z_r(t_i)) \right)^{-1}
\]

- **Pseudo-likelihood Breslow estimate**

\[
\hat{\Lambda}(t) = \sum_{i : t_i \leq t} p_i^{-1} \left( \sum_{r \in R_i} p_r^{-1}(\exp (\beta^T z_r(t_i))) \right)^{-1}
\]

- **Distortion factor Breslow estimate**

\[
\hat{\Lambda}(t) = \sum_{i : t_i \leq t} \left( \sum_{r \in R_i} \rho_r^{-1}(\exp (\beta^T z_r(t_i))) \right)^{-1}
\]