Analysis of Breast Cancer Survival Data with missing information on stage of disease and cause of death

Rino Bellocco\textsuperscript{1}, Nicola Orsini \textsuperscript{2}

\textsuperscript{1} Department of Medical Epidemiology and Biostatistics, Karolinska Institutet
\textsuperscript{2} Institute of Environmental Medicine, Karolinska Institutet

\textsuperscript{1}th Italian Stata User Group Meeting, October 25, 2004; Rome
Epidemiological findings indicate that breast cancer survival is related to socioeconomic factors. Women of lower socioeconomic status have generally been found to have poorer survival.

Epidemiological findings indicate that both breast cancer incidence and survival are related to socioeconomic factors. Women of lower socioeconomic status are at lower risk of developing breast cancer but tend to have poorer survival compared to socioeconomically more favored women.
A common problem in analysis of survival data is the presence of competing risks. When the cause of death is known, it is possible to study the effect of the exposure on cause-specific hazards by treating the deaths from other causes as censored observations in a Cox regression model.

As the follow-up increase, the time available for quality checking of the death certificates decreases and therefore the statistician has to face the dilemma whether to censor the data at an earlier period of time, where complete information on the endpoint is fully available, or to try using all the data by imputing the missing value of cause of death.

Furthermore, even if complete information on the main risk factor (social-economic status) is present, it is possible that some patient’s characteristics, such as tumor stage, might be missing for a particular reporting center.
Study Design: Cohort

- Linkage between the following Swedish population-based registers: the Cancer Register, five Regional Cancer Registers, the 1970, 1980, 1985 and 1990 Census databases, the Fertility Register, Emigration Register, and Cause of Death Register was made possible by using the individually unique National Registration Number (NRN) assigned to each resident.

- A total of 4645 women were diagnosed with invasive breast cancer as first diagnosis from January 1 to December 31 in Sweden in 1993. Of these, 1646 (35%) women have died as of December 31, 2001, the end of the follow-up period. However, 298 women died after December 31, 1998, the date after which the cause of death was unknown. The total number of women with ascertained cause of death was 1348, and 772 of these deaths (57.3%) were due to breast cancer.
Methods

• Standard survival analyzes are performed: the survival distribution is estimated by the Kaplan-Meier technique, and log-rank test is used to assess the influence of the main exposure variable.

• Cox proportional hazards regression model is fitted to the data to study how the estimates change according the different scenario of missing data for the covariates.
. stset ftime, fail(fail) id(lopnr) origin(entry) scale(365.4)

. sts graph if newsesw!=2, by(newsesw) failure ///
   xtitle("Years since Diagnosis") ///
   title("Woman Socio-Economic Status") ///
   ylabel(0 0.1 0.2) xlabel(0(1)8)

. sts test newsesw if newsesw!=2

. stcox newsesw if newsesw!=2
Partial Follow-up
. sts test newsesw if newsesw!=2

Log-rank test for equality of survivor functions

<table>
<thead>
<tr>
<th></th>
<th>Events</th>
<th>Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>newsesw</td>
<td>observed</td>
<td>expected</td>
</tr>
<tr>
<td>Low</td>
<td>299</td>
<td>273.97</td>
</tr>
<tr>
<td>High</td>
<td>125</td>
<td>150.03</td>
</tr>
<tr>
<td>Total</td>
<td>424</td>
<td>424.00</td>
</tr>
</tbody>
</table>

\[
\chi^2(1) = 6.46 \\
\text{Pr}>\chi^2 = 0.0110
\]
. stcox newsesw if newsesw!=2, nolog

Cox regression -- Breslow method for ties

No. of subjects = 2840 Number of obs = 2840
No. of failures = 424
Time at risk = 14069.12151
LR chi2(1) = 6.65
Log likelihood = -3312.1663 Prob > chi2 = 0.0099

| _t | Haz. Ratio | Std. Err. | z   | P>|z| | [95% Conf. Interval] |
|----|------------|-----------|-----|-----|----------------------|
| newsesw | 0.7634345 | 0.0813169 | -2.53 | 0.011 | 0.6195928 .9406697 |

9
Partial Follow-up, adjusted for stage
```
. stcox newsesw if newsesw!=2, strata(stage) nolog

Stratified Cox regr. -- Breslow method for ties

|                | Haz. Ratio | Std. Err. | z    | P>|z|  | [95% Conf. Interval] |
|----------------|------------|-----------|------|------|----------------------|
| newsesw        | 0.7934814  | 0.1026175 | -1.79| 0.074| 0.6158211 1.022395   |

Stratified by stage
```
Complete follow-up time, overall mortality
Multiple Imputation of missing cause of death can be done in different ways

- A logistic regression model can be fitted, in which for a woman with known cause of death the logit of the probability of dying of breast cancer is modeled as a function of complete observed covariates (marital status, age at diagnosis, income level).

- The second step, for a woman with missing cause of death is to generate a binary random variable with mean given my the fitted probability, repeating this \( m \) times
MICE Imputation of Cause of death

```
.tab type, missing

<table>
<thead>
<tr>
<th>type</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Die of OTHER</td>
<td>576</td>
<td>12.40</td>
<td>12.40</td>
</tr>
<tr>
<td>Die of BC</td>
<td>772</td>
<td>16.62</td>
<td>29.02</td>
</tr>
<tr>
<td>Alive</td>
<td>2,999</td>
<td>64.56</td>
<td>93.58</td>
</tr>
<tr>
<td>.</td>
<td>298</td>
<td>6.42</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Total | 4,645 | 100.00 |
```

forvalues i = 1(1)100 {
    uvis logit type marstat newageb* incgrb* if type != 2,gen(bmiss'i')
}
uvis imputes type from marstat, newage, incgr according to the following algorithm (van Buuren et al. (1999) for further technical details):

- Estimate the vector of coefficients (beta) by regressing the nonmissing values of type on marstat, newage, incgr. Predict the fitted values of the logit of the probability of type = 1 at the nonmissing observations of type.

- Draw at random a value (sigma*) from the posterior distribution of the residual standard deviation.

- Draw at random a value (beta*) from the posterior distribution of beta, allowing, through sigma*, for uncertainty in beta.
Use beta* to predict the fitted values of the logit of the probability of type = 1 at the missing observations of type.

(Prediction matching) For each missing observation of type with prediction given by the step above, find the nonmissing observation of type whose prediction given by the step 1 on observed data is closest to the fitted values. This closest nonmissing observation is used to impute the missing value of type.
**Imputation results**

```
. summarize propfail

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>propfail</td>
<td>100</td>
<td>.5764763</td>
<td>.0192317</td>
<td>.5236938</td>
<td>.6160389</td>
</tr>
</tbody>
</table>
```
Imputed Kaplan-Meier Survival distribution

Woman Socio–Economic Status

Follow-up time

01jan1993 01jul1994 01jan1996 01jul1997 01jan1999 01jul2000

F(t)
tests the equality of the survivor function across socio-economic status
Stage of disease imputation

- Next step will be to model missingness in stage of disease.
- Missingness only depends on data not reported by one of the region of the cancer register.
- We will adjust the effect of social status by stage and age under the assumption that the stage distribution condition to social status is similar to the one of the other reporting regions.
- Multiple imputation will be performed also in this case.
### Imputation results

```
.tab fail50

<table>
<thead>
<tr>
<th>fail50</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3,688</td>
<td>79.40</td>
<td>79.40</td>
</tr>
<tr>
<td>1</td>
<td>957</td>
<td>20.60</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>4,645</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>
```
<table>
<thead>
<tr>
<th>Tumor Size</th>
<th>Low</th>
<th>High</th>
<th>Not Emp</th>
<th>.</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>870</td>
<td>447</td>
<td>546</td>
<td>71</td>
<td>1,934</td>
</tr>
<tr>
<td></td>
<td>46.98</td>
<td>45.24</td>
<td>34.06</td>
<td>27.20</td>
<td>41.11</td>
</tr>
<tr>
<td>2</td>
<td>428</td>
<td>222</td>
<td>456</td>
<td>44</td>
<td>1,150</td>
</tr>
<tr>
<td></td>
<td>23.11</td>
<td>22.47</td>
<td>28.45</td>
<td>16.86</td>
<td>24.45</td>
</tr>
<tr>
<td>3</td>
<td>46</td>
<td>21</td>
<td>48</td>
<td>8</td>
<td>123</td>
</tr>
<tr>
<td></td>
<td>2.48</td>
<td>2.13</td>
<td>2.99</td>
<td>3.07</td>
<td>2.61</td>
</tr>
<tr>
<td>4</td>
<td>17</td>
<td>5</td>
<td>33</td>
<td>1</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>0.92</td>
<td>0.51</td>
<td>2.06</td>
<td>0.38</td>
<td>1.19</td>
</tr>
<tr>
<td>.</td>
<td>491</td>
<td>293</td>
<td>520</td>
<td>137</td>
<td>1,441</td>
</tr>
<tr>
<td></td>
<td>26.51</td>
<td>29.66</td>
<td>32.44</td>
<td>52.49</td>
<td>30.63</td>
</tr>
<tr>
<td>Total</td>
<td>1,852</td>
<td>988</td>
<td>1,603</td>
<td>261</td>
<td>4,704</td>
</tr>
<tr>
<td></td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>
Imputation results

mvis stage  reg2 reg3 reg4 reg5 newage  _d lnt ///
using breast, m(5)  genmiss(m_)

Imputation results

```
. micombine stcox newsesw if newsesw!=2, strata(stage) eform(exp)
version = 8.2

Multiple imputation parameter estimates (5 imputations)
-----------------------------------------------------------------
     _t | exp      Std. Err.     z   P>|z|   [95% Conf. Interval]
--------+--------------------------------------------------------
   newsesw | 0.7284751  0.0746985  -3.09  0.002   0.5958429   0.8906307
-----------------------------------------------------------------
2840 observations.
```
References


