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

Rino Bellocco[1], Nicola Orsini [2]

[1] Department of Medical Epidemiology and Biostatistics , Karolinska Institutet

[2] Institute of Environmental Medicine, Karolinska Institutet

Ith Italian Stata User Group Meeting , October 25, 2004; Rome

Introduction

- 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.

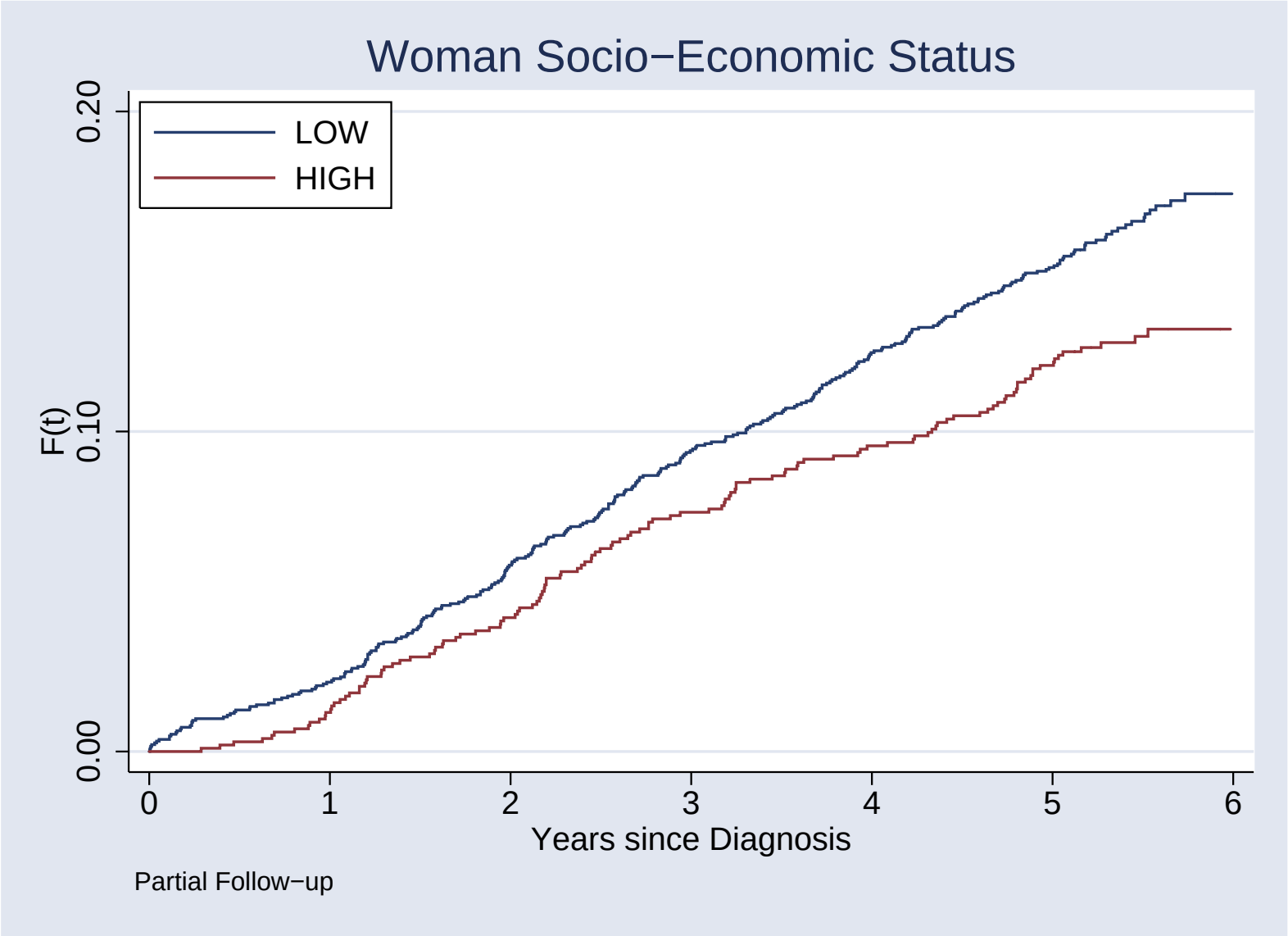
Study Design: Cohort

- 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

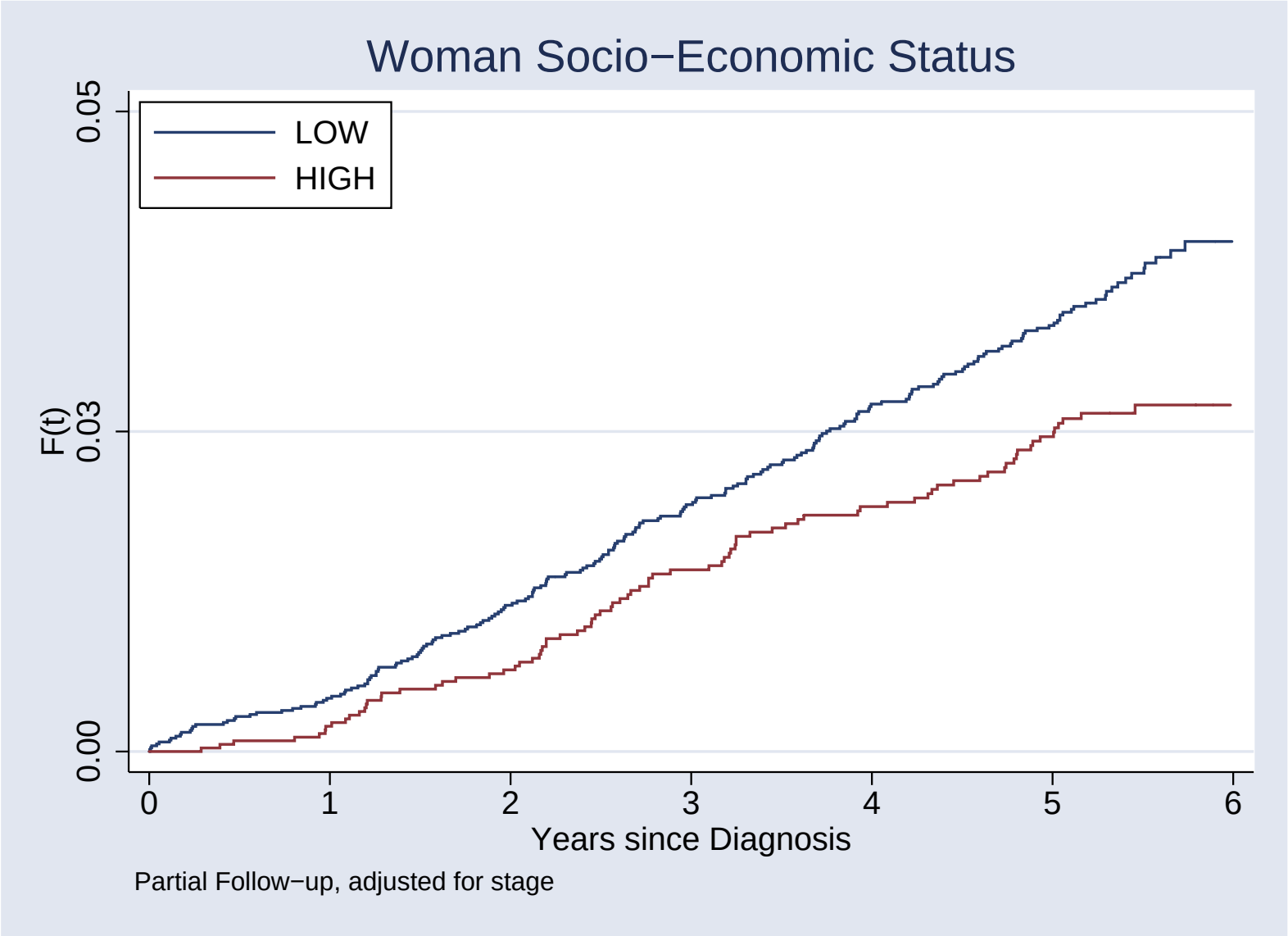
- Standard survival analyzes are performed: the survival distribution is estimated by the Kaplan-Meier technique, and log-rank test is used to assess the influence of the main exposure variable.
- Cox proportional hazards regression model is fitted to the data to study how the estimates change according 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
```

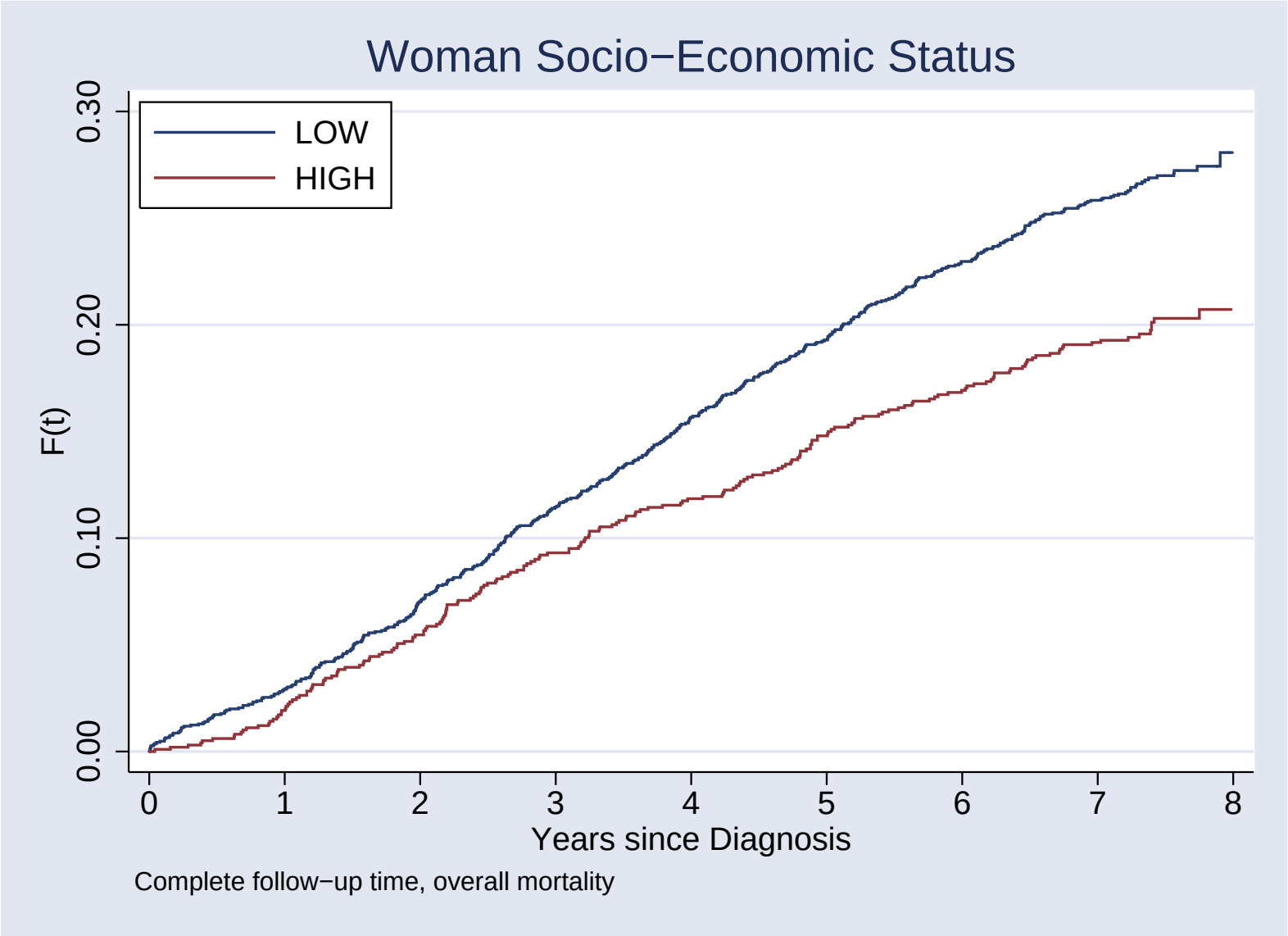

Woman Socio−Economic Status
LOW
HIGH
F(t)
0.00
0.10
0.20
0
1
2
3
4
5
6
Years since Diagnosis
Partial Follow−up

```
. sts test newsesw if newsesw!=2


Log-rank test for equality of survivor functions
         |   Events          Events
newsesw  |  observed        expected
---------+-------------------------
Low      |       299          273.97
High     |       125          150.03
---------+-------------------------
Total    |       424          424.00

                chi2(1) =       6.46
                Pr>chi2 =     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 |   .7634345   .0813169    -2.53   0.011     .6195928    .9406697
------------------------------------------------------------------------------
```


Woman Socio−Economic Status
LOW
HIGH
F(t)
0.00
0.03
0.05
0
1
2
3
4
5
6
Years since Diagnosis
Partial Follow−up, adjusted for stage

```
.  stcox newsesw if newsesw!=2, strata(stage) nolog

Stratified Cox regr. -- Breslow method for ties

No. of subjects =         2840   Number of obs   =      2056
No. of failures =          424
Time at risk    =  14069.12151
                                 LR chi2(1)      =      3.30
Log likelihood  =   -1796.0402   Prob > chi2     =    0.0693

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     newsesw |   .7934814   .1026175    -1.79   0.074     .6158211    1.022395
------------------------------------------------------------------------------
                                                      Stratified by stage
```


Woman Socio−Economic Status
LOW
HIGH
F(t)
0.00
0.10
0.20
0.30
0
1
2
3
4
5
6
7
8
Years since Diagnosis
Complete follow−up time, overall mortality

Imputation of Cause of death

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

                type |      Freq.     Percent        Cum.
---------------------+-----------------------------------
        Die of OTHER |        576       12.40       12.40
           Die of BC |        772       16.62       29.02
               Alive |      2,999       64.56       93.58
                   . |        298        6.42      100.00
---------------------+-----------------------------------
               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):

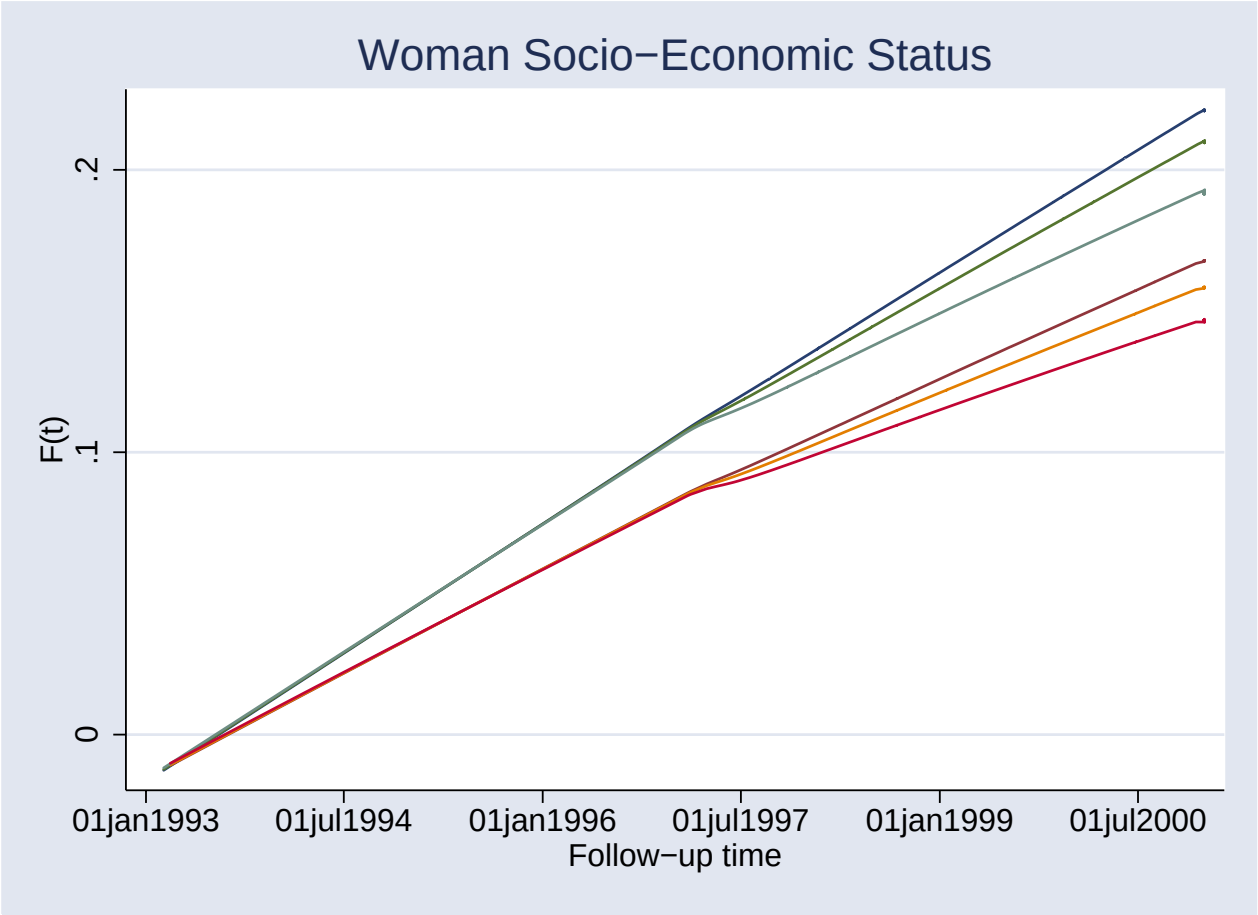
- Estimate the vector of coefficients (beta) by regressing the nonmissing values of *type* on *marstat, newage, incgr*. Predict the fitted values of the logit of the probability of type = 1 at the nonmissing observations of *type*.
- Draw at random a value (sigma*) from the posterior distribution of the residual standard deviation.
- Draw at random a value (beta*) from the posterior distribution of beta, allowing, through sigma*, for uncertainty in beta.

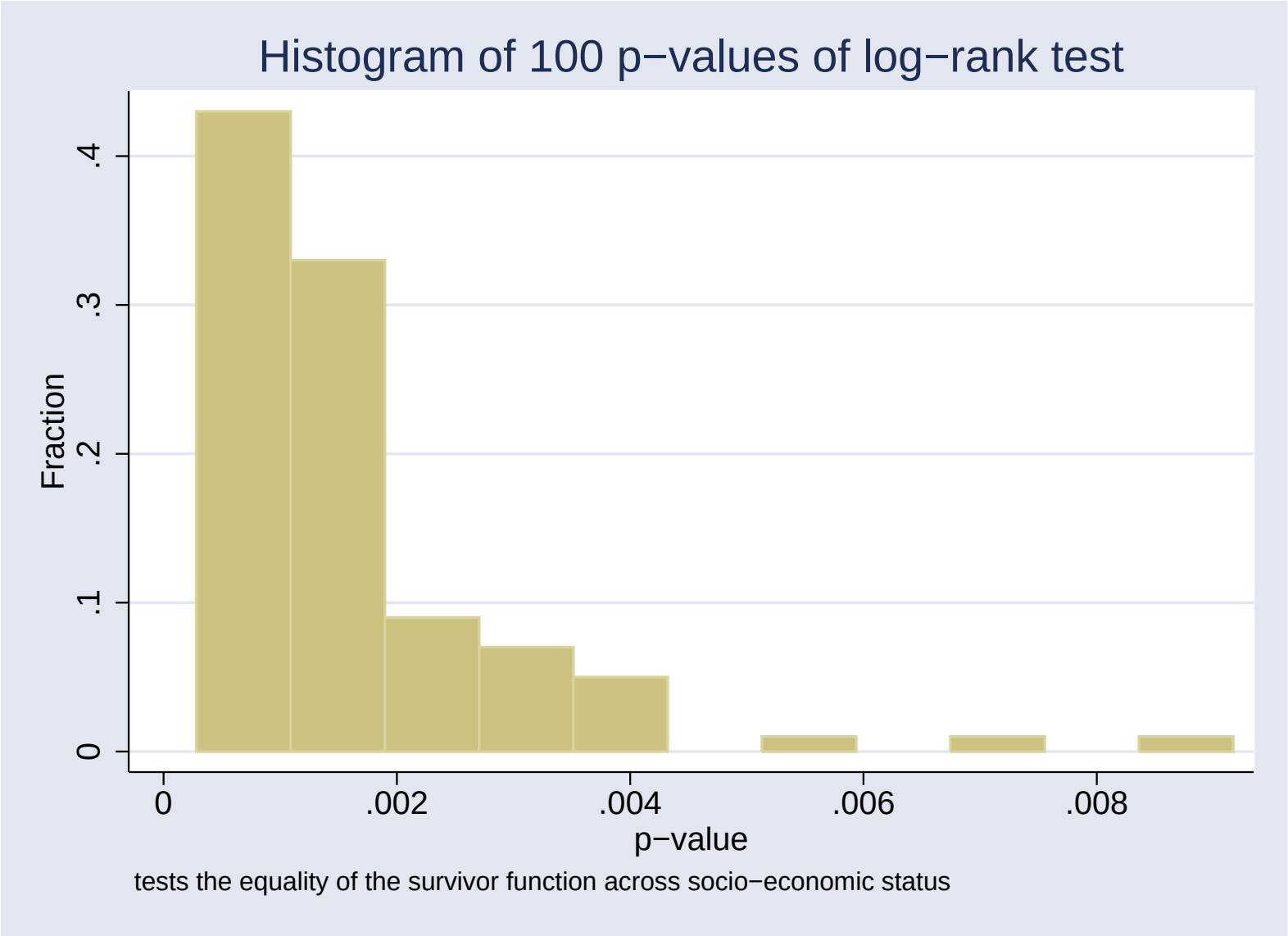
- Use beta* to predict the fitted values of the logit of the probability of type = 1 at the missing observations of *type*.
- (Prediction matching) For each missing observation of *type* with prediction given by the step above, find the nonmissing observation of *type* whose prediction given by the step 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

 Variable |   Obs        Mean    Std. Dev.       Min        Max
----------+--------------------------------------------------------
 propfail |   100    .5764763    .0192317   .5236938   .6160389
```

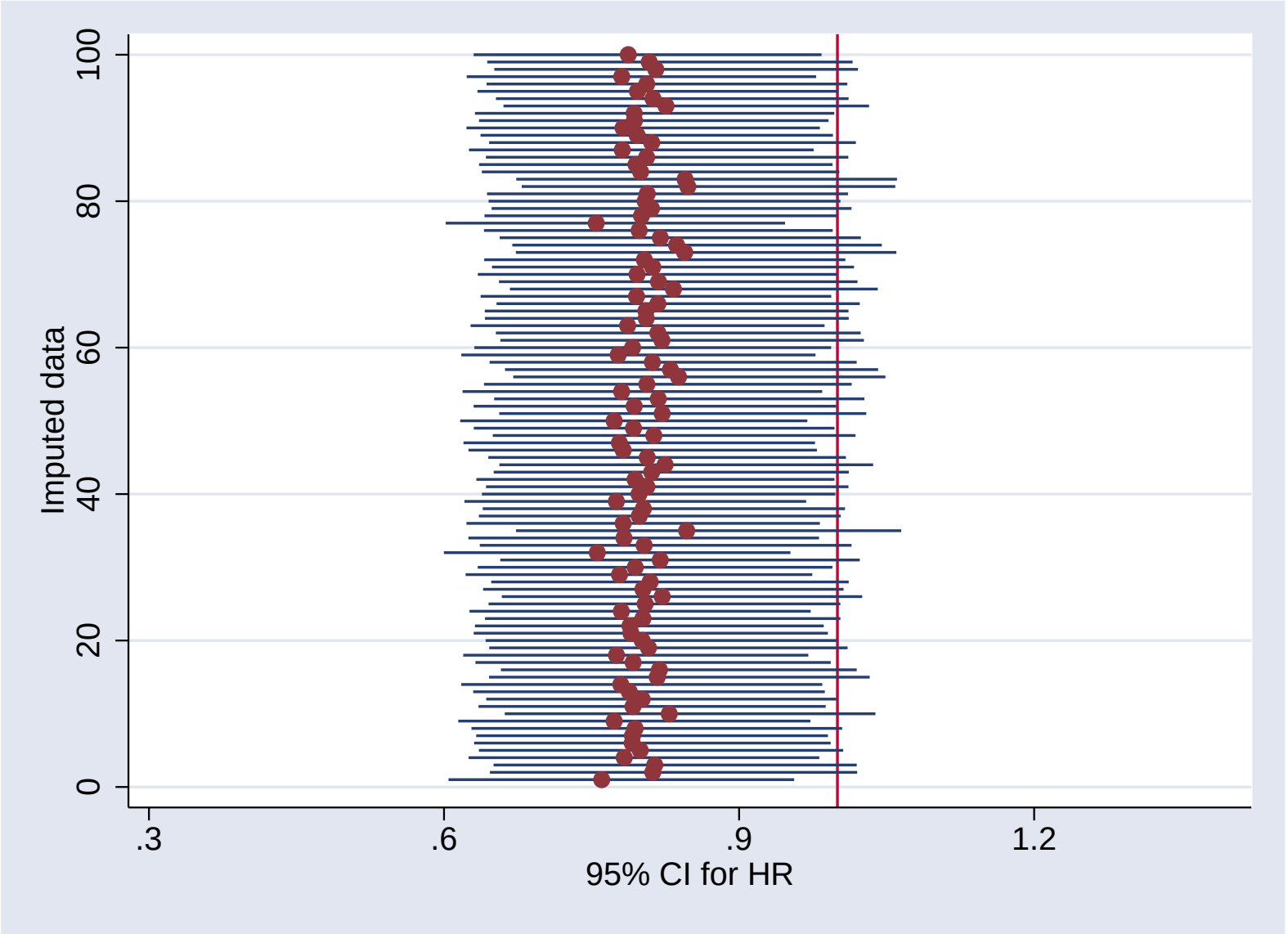
Imputed Kaplan-Meier Survival distribution




Histogram of 100 p−values of log−rank test
Fraction
0
.1
.2
.3
.4
0
.002
.004
.006
.008
p−value
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.


Imputed data
0
20
40
60
80
100
.3
.6
.9
1.2
95% CI for HR

Imputation results

```
. tab fail50

     fail50 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      3,688       79.40       79.40
          1 |        957       20.60      100.00
------------+-----------------------------------
      Total |      4,645      100.00
```

	New SEI woman				
Tumor Size	Low	High	Not Emp	.	Total
1	870	447	546	71	1,934
	46.98	45.24	34.06	27.20	41.11
2	428	222	456	44	1,150
	23.11	22.47	28.45	16.86	24.45
3	46	21	48	8	123
	2.48	2.13	2.99	3.07	2.61
4	17	5	33	1	56
	0.92	0.51	2.06	0.38	1.19
.	491	293	520	137	1,441
	26.51	29.66	32.44	52.49	30.63
Total	1,852	988	1,603	261	4,704
	100.00	100.00	100.00	100.00	100.00

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 |.7284751   .0746985    -3.09   0.002     .5958429    .8906307
------------------------------------------------------------------------------
2840 observations.
```

References

Andersen, J., Goetghebeur, E., Ryan, L.. (1996). *Missing Cause of death information in the analysis of survival Data*. **15**, 2191-2201.

Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman and Hall: London.

Faggiano, F., Partanen, T., Kogevinas, M., Boffetta, P. (1997). Socioeconomic differences in cancer incidence and mortality. *IARC Scientific Publications*, **138**, 65-176.

Garne, J.P., Aspegren, K., Moller, T. (1995). *Validity of breast cancer registration from one hospital into the Swedish National Cancer Registry 1971-1991*. Acta Oncologica, 34(2):153-6.

Geenland, S., Finke, W.D. (1995). *A critical look at methods for handling missing covariates in epidemiologic regression analysis*, **142(12)**, 1255-1264.

Vågerö , D., Persson, G. (1987). Cancer survival and social class in Sweden. *Journal of Epidemiology and Community Health.*, **41(3)**, 204-9.

National Board of Health and Welfare (1996). Cancer incidence in Sweden 1993. *Centre for Epidemiology, National Board of Health and Welfare*. Stockholm, Sweden.

van Buuren, S., Boshuizen, H.C., and Knook, D.L. (1999). Multiplicative imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*. **18**: 681-694.