Computer Practical Day III
Exercises

1. **Larynx carcinoma patients**

   Consider the following figure (borrowed from a published article). It compares the survival of larynx carcinoma patients with and without a secondary tumor. Time is measured since detection of the first tumor.

   a. Do you see anything counter-intuitive in the graph?
   b. How can you correctly model the effect of the second tumor on survival?

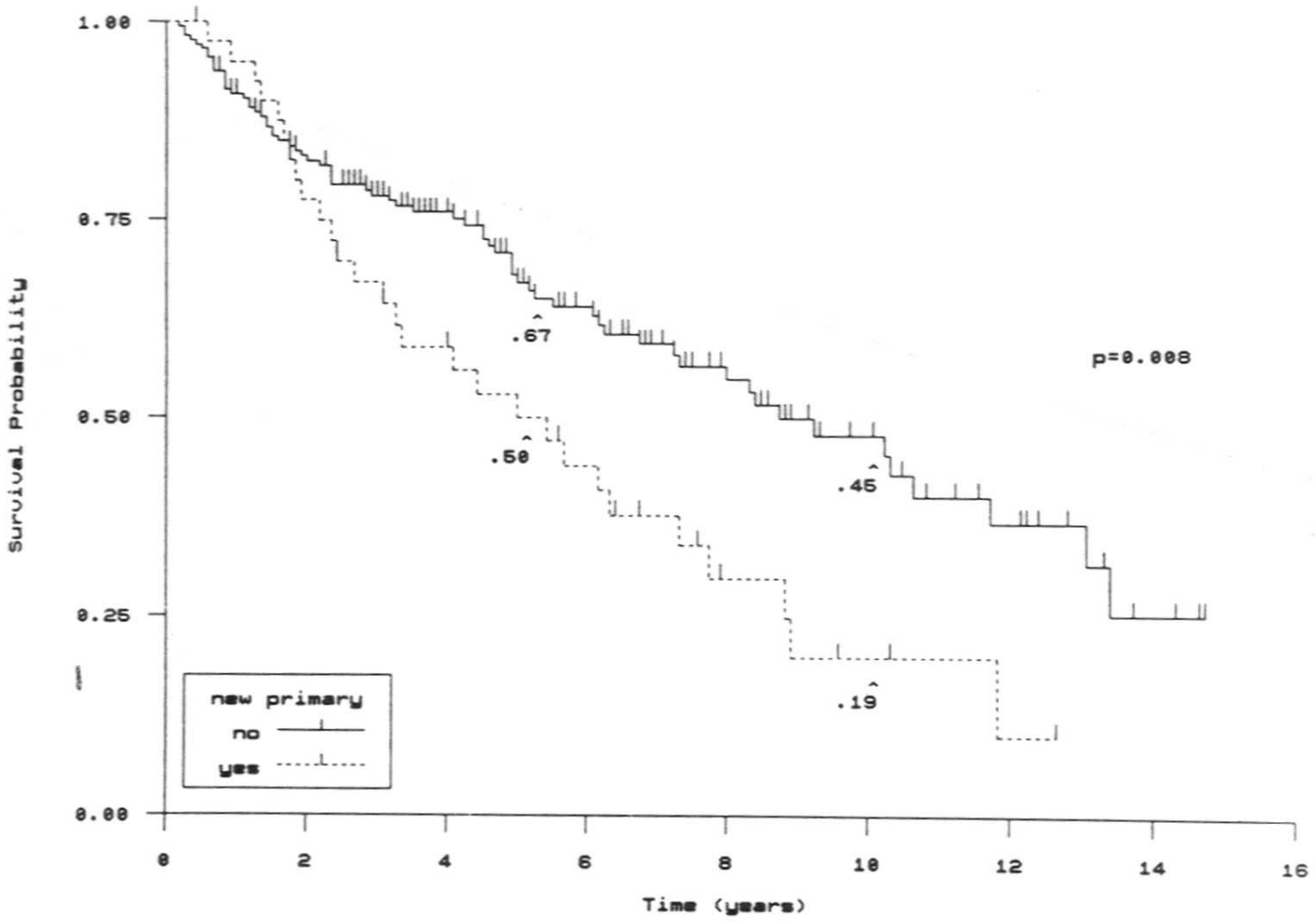

Fig. 3. The actuarial survival for patients who did (41) or did not (177) develop an SMN.

2. **Ovarian cancer data**

We consider data on 358 patients with ovarian carcinoma, all of them treated with comparable chemotherapy. The follow-up for each patient is at least 4 years since the start of the treatment. As explanatory variables we use: KARN, DIAM and FIGO. KARN indicates the Karnovsky-index at the start of the treatment. It measures the mobility of the patient, the lower the worse the condition. DIAM indicates the residual tumour after operation. FIGO status indicates the growth of the tumour. Today we look at the probability of survival as a function of time. The data are in the file **OVA.SAV**. There you can find the coding of the explanatory variables. We use the variables

time		survival time in days
death	1	if the patient died
	0	if the patient is censored

Let's see how we can fit a COX model to these data.

a. Start by making a Kaplan-Meier curve for both FIGO groups separately. Also make a plot of the cumulative hazards and check (just by looking at the graphs) for proportionality. Try to get a rough estimate for the relative risk of FIGO=1 compared to FIGO=0. Compute the log-rank test and write down the test statistic and the p-value. All this can be found under Analyze, Survival, Kaplan-Meier. Take FIGO as factor. Unselect the survival tables under statistics to avoid an overload of output.

b. Execute a Cox regression with FIGO as covariate. Choose the command Analyze; Survival; Cox regression. Look at the estimator for beta in the Cox-model. Compute the relative risk (hazard ratio) on death in group FIGO=1 compared to FIGO=0. Is there a significant difference between the two groups? Compare the value of the log rank test statistic and the one associated with the coefficient in the model.

c. Compute from the data in the output manually the 95% confidence interval for the relative risk of death (the hazard ratio) for FIGO=1 versus FIGO=0. Check yourself by rerunning the Cox regression with the appropriate options.

d. Again run the Cox regression but define FIGO as a categorical variable. What has changed in the output and why?

e. Repeat this but now plot the survival curves in both groups using the "separate lines" option under the Plot topic, still under Cox regression. What is the difference with the Kaplan-Meiers from part (a)?

f. Carry out Cox regressions for KARN and DIAM separately as categorical variables. Make graphs of the model curves. Provide an argument why these 2 variables may be entered into the model as continuous rather than categorical variables.

g. Now enter KARN, FIGO and DIAM as covariates into the model. Check the output and compare the coefficient of FIGO with the value found earlier. What is now your

interpretation of the coefficient of FIGO?

h. Check whether inclusion of interaction terms could improve the model fit.
Save in a model you deem correct, the linear predictor XBETA (write down how it is defined).

i. Subdivide the data in three roughly equal groups on the basis of XBETA (make SPSS compute percentiles for you). In this way you get a subgroup with a good; one with a moderate and one with a bad prognosis.

j. Ask SPSS to generate the Kaplan Meier for these three subgroups. Do you think you can make an acceptable prediction of survival using these data and models?

k. In the new model, predict survival for two patients, whose only difference is their FIGO score (take the mean value for the other covariates).

3. **Time-dependent covariates**

The bmtsmall.sav SPSS data file consists of 137 leukemia patients that have received a bone-marrow transplantation. It contains the following variables:

group
: Disease Group 1-ALL, 2-AML Low Risk, 3-AML High Risk

dfsdays
: Disease Free Survival Time (Time To Relapse, Death Or End Of Study)

dfsstat
: Disease Free Survival Indicator 1-Dead Or Relapsed, 0-Alive Disease Free)

tp
: Time To Platelet Recovery

dp
: Platelet Recovery Indicator 1-Platelets Returned To Normal, 0-Platelets Never Returned to Normal

For this exercise, data about other intermediate events and about characteristics at baseline have been removed.

The aim of this exercise is to examine the influence of intermediate events on disease-free survival, and to see how the presence of such an intermediate event may influence the effect of other fixed covariates in a Cox regression model. As an example of such an intermediate event, we consider the return of the patient's platelet count to a self-sustaining level (platelet recovery).

a. Make a Kaplan-Meier for disease-free survival (DFS) for the three risk groups (variable 'group'). Test whether DFS differs between these three groups.

b. Fit a Cox regression model with group as sole covariate. Make sure group is considered as a categorical covariate; take the ALL risk group as reference category. Ask for 95% confidence intervals of the hazard ratios (under Options). Interpret the results. Do they agree with the Kaplan-Meier plot?

c. Now we include as time-dependent covariate the occurrence of platelet recovery. In this way we can quantify how the occurrence of a platelet recovery influences DFS for these patients. The time-dependent covariate $Z_P(t)$ is defined as follows

$Z_P(t) = \begin{cases} 0, t \le T_P \\ 1, t > T_P \end{cases}$, where T_P is the time of platelet recovery.

Click Analyze -> Survival -> Cox w/ Time-Dep Cov for a Cox regression analysis with time-dependent covariates.

You first get a window in which you can define the time-dependent covariate. This time-dependent covariate will be called T_COV_ in the Cox regression analysis. We type in a formula that defines $Z_P(t)$ as above. One way to do this is as 1*(T_ > tp), which will be 1 if T_ (time) is greater than the time of platelet recovery (tp) and 0 otherwise.

Click on Model in the upper right part of the window. You will get the by now familiar window from Cox regression, the only exception being the presence of the time-dependent covariate T_COV_ in the list of covariates. Select group (using the same coding as before) and T_COV_ as covariates in the model.

Ask for 95% confidence intervals again (under Options), and click OK.

Study the results. What is the effect of the time-dependent covariate? How should we interpret this? Have the effects of the risk groups changed?

d. We could also (incorrectly!!) use platelet recovery as if it was known at baseline, by simply adding dp to the Cox regression model with group as covariate. This will create two groups, one of patients who will have a platelet recovery during the follow-up (dp=1) and one of patients who will not (dp=0). Perform this analysis. Why is it wrong? Is the bias in the direction you would expect?