A Practical Guide on Modeling Competing Risk Data

Prepared by

Giorgos Bakoyannis*

And

Giota Touloumi*

on behalf of CASCADE collaboration

*Dept. of Hygiene, Epidemiology and Medical Statistics, Medical School, University of Athens

Citation: A Practical Guide on Modeling Competing Risk Data by Giorgos Bakoyannis and Giota Touloumi on behalf of CASCADE Collaboration. www.cascade-collaboration.org downloaded on [insert date of download].

Acknowledgements: Many thanks to all CASCADE collaborators and particularly to Sarah Walker and Ronald Geskus for their helpful comments.

Table of contents

1. Introduction…… 2
1.1 Theoretical approaches…… 2
1.1.1 Latent failure times…… 2
1.1.2 Joint distribution of time and cause of failure…… 3
1.2 Estimable and non-identifiable quantities…… 3

2 Statistical modeling…….. 4
2.1 Cause-specific hazard…… 5
2.1.1 Implementation of the methods using STATA…….. 7
2.1.2 Implementation of the methods using R…… 11
2.2 Cumulative incidence…… 13
2.2.1 Implementation of the methods using STATA…….. 15
2.2.2 Implementation of the methods using R…… 17

3 Conclusions..…… 21

Appendix 1. Description of the dataset…….. 22
Appendix 2. R software and package downloads……24
Appendix 3. Comparison of the methods……... 27

References……... 34

1. Introduction

In cohort studies and clinical trials, time to an event is being frequently studied. Competing risk data are encountered when subjects under study are at risk of more than one mutually exclusive events, like death from different causes. The term competing risks also refers to data where the different possible events are not mutually exclusive but the interest lies on the first coming event. For example, in HIV-1 seropositive subjects receiving highly active antiretroviral therapy (HAART), treatment interruption (TI) and initiation of a new HAART regimen act as competing risks for the 1st major change in HAART.

Within the CASCADE an analysis to identify prognostic factors for TIs, treating shift to a new HAART regimen as a competing risk under the scenario of administrative loss to follow-up has been done and results have been published (Touloumi et al, 2006). Based on this work, we proposed to re-analyze the same dataset with the aim to apply and compare the various existing modeling techniques for analyzing such data. This report is based on the results of this exercise.

The basic theoretical approaches used for competing risks along with the basic semi-parametric models are briefly presented. The different methods are illustrated using data from CASCADE. More specifically, we are interested in modelling time and type of 1st major change after HAART initiation (i.e. TI or shift to a new HAART regimen). The data used were pooled in 2003 and contained information on 8300 seroconverters. The inclusion criteria were: a) Initiation of stable 1st HAART (duration >=90 days) at least one year after seroconversion; b) availability of CD4 cell count and HIV-RNA measurements at and during HAART. For simplicity only gender and mode of infection have been considered as covariates in all examples presented here.

1.1 Theoretical approaches

Competing risk models can be defined and described in two alternative ways. The first one relies on k hypothetical failure times, one for each possible cause of failure, whereas the second is related to the joint distribution of time and cause of failure.

1.1.1 Latent failure times

The first approach to competing risks assumes the existence of k latent failure times, one for each possible type of failure. Let T_j be the time to failure from cause j. In studies where competing risks are present

we only observe the minimum of the latent failures times (T) and the corresponding cause of failure (C). More technically we observe:

$$T = \min_j \{T_j : j = 1,2,...,k\}$$
$$C = \arg\min_j \{T_j : j = 1,2,...,k\}$$

Function $\arg\min_j \{T_j\}$ gives c if $\min_j \{T_j\} = T_c$. The joint distribution of the latent failure times and the corresponding marginals can be defined by the multivariate and marginal survival functions respectively:

$$H(t_1,t_2,...,t_k) = \Pr(T_1 > t_1, T_2 > t_2, ..., T_k > t_k)$$
$$S_j^*(t_j) = \Pr(T_j > t_j) = H(0,0,...,t_j,...,0)$$

1.1.2 Joint distribution of time and cause of failure (multi-state approach)

The recent approach to competing risks is considering the joint distribution of failure time T and cause of failure C, two observable random variables. Basic quantity that derives from this approach is the cause-specific hazard function:

$$\lambda_j(t) = \lim_{h \to 0} \frac{\Pr(t \le T < t+h, C = j \mid T > t)}{h}, \quad \text{j=1,2,...,k}$$

Cause-specific hazard for a cause j is the instantaneous failure rate from this cause in the presence of all other possible causes of failure. The probability of failure from cause j until time t in the presence of all other possible causes is known as cause-specific cumulative incidence and depends on the cause-specific hazards for **all** other causes.

$$F_j(t) = \Pr(T \le t, C = j) = \int_0^t \lambda_j(u) \exp\left\{ -\int_0^u \sum_{c=1}^k \lambda_c(w)\, dw \right\} du, \quad \text{j=1,2,...,k}$$

1.2 Estimable and non-identifiable quantities

Not all quantities that are being involved in theoretical development of competing risks can be estimated from the data without

making further assumptions. The likelihood function for competing risk data has the following form:

$$L = \prod_{c=1}^{k} \prod_{i=1}^{n} \lambda_c(t_i)^{d_{ic}} \exp\left\{ -\int_0^{t_i} \lambda_c(u)\,du \right\} \qquad (1)$$

$$d_{ic} = I(C_i = c)$$

The above likelihood is a function of the cause-specific hazard functions. Thus, cause-specific hazards and functions of them (like the cumulative incidence) are estimable from the data. The joint and marginal distributions of latent failure times cannot be estimated from the data without making strong assumptions (like independence between latent failure times) that are not testable from the data.

2 Statistical modeling

First of all, it should be noted that one minus Kaplan-Meier estimate (based on cause-specific hazard) does not always provide an appropriate estimate for cumulative incidence for cause j as it generally overestimates that quantity. It can only be applied to estimate cumulative incidence for cause j in the hypothetical situation where failures from other causes have been eliminated and the cause-specific hazard of interest does not change after this elimination (i.e. failure times for the different events are independent) (Gaynor et al, 1993).

Statistical models for competing risks are mainly focalized on estimable quantities. It must be noted that the effect of a covariate on the cumulative incidence may be different from its effect on the cause-specific hazard function for the corresponding cause of failure. This is due to the fact that cumulative incidence is a function of the cause specific hazards for all the possible causes of failure (1) and so the effect of a covariate on the cumulative incidence for some cause does not depend only on the effect of the covariate on the cause-specific hazard for the corresponding cause, but also on the respective effect on the cause-specific hazard functions for all the other causes. To make it clearer, consider the following example: suppose a new drug reduces the instantaneous risk of an event A by 50%. It is used in two groups of patients - one with a high rate of event A and a low rate of another competing event B, and the second group with a high rate of B and a lower rate of event A. Because patients in this second group are at a much higher risk of B, the overall effect of the new drug on the cumulative incidence of event A will be smaller because patients are getting event B (not affected by the drug) first. However, in the first group the 50% lower

cause-specific hazard for event A will leave many more people at risk for event B - so the cumulative incidence of event B is likely to rise just because patients are left at risk for this event because the new drug is stopping them getting A. Consequently, there exist different modeling techniques for cause-specific hazard and cumulative incidence. In this report only the basic semi-parametric proportional hazard models are being considered because of their flexibility (no distributional assumptions on T) and the availability of software for fitting these models.

2.1 Cause-specific hazard

As noted above cause-specific hazard is the instantaneous failure rate from a specific cause in the presence of all the others. Estimation of cause-specific hazard model parameters can be undertaken by maximizing the likelihood function (1). Parameter estimates for the cause-specific hazard of a cause j can be obtained by maximizing the factor of the likelihood involving cause j (when the different factors have no common parameters). Moreover, it should be noted that this factor is the same with the likelihood function that would be obtained by treating failures from all other causes, except for j, as censored observations. For the cause specific-hazard functions we can assume a semiparametric proportional hazards model:

$$\lambda_j(t;x) = \lambda_{j0}(t)\exp\{\beta_j' x\}, \quad \text{j=1,2,...,k}$$

The corresponding partial likelihood function is:

$$L^p(\beta_1,\beta_2,...,\beta_k) = \prod_{c=1}^{k} L_c(\beta_c)$$

$$L_c(\beta_c) = \prod_{i=1}^{d_c}\left(\frac{\exp\{\beta_c' x_{ic}\}}{\sum_{l\in R(t_{ic})}\exp\{\beta_c' x_l\}}\right)$$

Parameters' estimates for cause-specific hazard for cause j can be obtained by maximizing the likelihood factor j. This can be applied by treating observations with failure from all other causes except of j as censored and fitting a Cox proportional hazards model on these data using (almost) any statistical package.

It is possible to fit simultaneously cause-specific hazard models for all causes and to test the equality of effects of specific covariates on

different failure types (i.e. $H_0 : \beta_{AIDS} = \beta_{death}$ in the model $\lambda_j(t;age) = \lambda_{j0}(t)\exp\{\beta_j \times age\}$, j=AIDS, death), through a data augmentation method. This method can be applied in any statistical package fitting Cox proportional hazards model. To do that, one needs to multiply the records k times, one for each possible failure type, and to generate a failure type (ft) identifier so that each record of a subject corresponds to one cause of failure. The failure indicator takes the value 1 in the record of the subject that corresponds to the actual failure cause and 0 in the remaining records of this subject. Failure indictor takes the value 0 in all corresponding records for subjects that have not failed (censored). For example, in a setting with k competing risks, the records of a subject i who has failed from cause c at time t_i should be:

id	failure type	failure indicator	failure time
i	1	0	t_i
i	2	0	t_i
.	.	.	.
.	.	.	.
.	.	.	.
i	c-1	0	t_i
i	c	1	t_i
i	c+1	0	t_i
.	.	.	.
.	.	.	.
.	.	.	.
i	k	0	t_i

There are two versions of the model for the cause-specific hazards: one is stratified by failure type and the other includes failure mode as a covariate (assuming proportional baseline hazards for the different failure types). If the baseline hazards for the different failure types are really proportional, then the "unstratified" method of analysis is more efficient than the "stratified". In both versions of the method, interaction terms of the covariates with failure type are included in the model. In this way the effects of the covariates on different failure types are not constrained to be equal. The test for the equality of the effect of a covariate on cause-specific hazard for different causes of failure is the test of the appropriate interaction term. In order to gain efficiency we can omit non-significant interaction terms and treat the effects of the corresponding covariates as equal across different failure modes. Robust estimates for the parameters' variances can be used to account for the correlation caused by the

multiplication of the observations and for possible model misspecification. It has been pointed out (Putter et al, 2007) that it is not necessary to account for the correlation of the multiplied observations when each subject has at most one event, but robust estimates can still be used to correct for other types of model misspecification.

In summary, one could fit: **a)** a semi-parametric proportional hazards model for cause-specific hazard function of a single type of failure or **b)** a semi-parametric proportional hazards model for cause-specific hazard functions of all types of failure simultaneously which allows testing the equality of effects of specific covariates on cause-specific hazard functions of different failure types. The latter model can be fitted as: **i)** a stratified by failure type model (which does not assume proportionality between different causes of failure) and **ii)** a model containing failure type as a covariate assuming proportional hazards for the different causes of failure.

2.1.1 Implementation of the methods using STATA

The data used to apply the above described methods are presented in Appendix 1. The data for some selected patients are shown below.

	id	time	event	gender	expo
20.	20	1.913758	new HAART	M	MSM
21.	21	.807666	new HAART	M	MSM
22.	22	3.917865	Still on HAART	M	MSM
23.	23	5.229295	Still on HAART	M	MSM
24.	24	3.088296	Still on HAART	F	MSW
25.	25	2.20397	TI	M	MSM
26.	26	.9691992	Still on HAART	M	MSW
27.	27	3.364819	Still on HAART	M	MSM
28.	28	1.21013	Still on HAART	M	MSM
29.	29	.3641342	TI	M	MSM
30.	30	.936345	TI	M	MSM

a) To fit a semi-parametric proportional hazards model for cause-specific hazard of treatment interruption (TI) we define the failure indicator to take 1 in observations with TI and 0 otherwise (i.e. shift to a new HAART regimen or censoring) and we declare the data as survival data:

```
use changeHAART, clear
```

```
gen fail=(event==1) /* defines the failure
indicator*/

stset time, f(fail) /* declares the data as
survival */
```

or

```
stset time, f(event==1)
```

After executing the above commands we fit the model as usual:

```
xi: stcox i.gender i.expo /* fits the Cox-
proportional hazards model with covariates gender
and mode of infection */
```

If we want to obtain the estimated coefficients (i.e. $\log_e$(estimated HR)) we include `nohr` option:

```
xi: stcox i.gender i.expo, nohr
```

It should be noted that the default method for handling ties is Breslow's method. To use Efron's approximation (which is preferable) we should add the option `efron`:

```
xi: stcox i.gender i.expo, nohr efron
```

b) To fit the model for TI and new HAART initiation simultaneously we have first to duplicate (as we have 2 events) the observations and then to generate failure type identifier:

```
use changeHAART, clear

expand 2 /* multiplies the records as many times
as the number of possible failure types */

sort id

by id: gen ft=_n /* the first record of each
subject corresponds to TI (1) and the second to
new HAART initiation (2) */
```

Then we generate the failure indicator which takes value "1" in records with failures

```
gen fail=(event==ft) /* assuming that there are
no missing values in "fail" */
```

The augmented data for the selected patients are shown below.

	id	time	event	ft	fail	gender	expo
39.	20	1.913758	new HAART	1	0	M	MSM
40.	20	1.913758	new HAART	2	1	M	MSM
41.	21	.807666	new HAART	1	0	M	MSM
42.	21	.807666	new HAART	2	1	M	MSM
43.	22	3.917865	Still on HAART	1	0	M	MSM
44.	22	3.917865	Still on HAART	2	0	M	MSM
45.	23	5.229295	Still on HAART	1	0	M	MSM
46.	23	5.229295	Still on HAART	2	0	M	MSM
47.	24	3.088296	Still on HAART	1	0	F	MSW
48.	24	3.088296	Still on HAART	2	0	F	MSW
49.	25	2.20397	TI	1	1	M	MSM
50.	25	2.20397	TI	2	0	M	MSM
51.	26	.9691992	Still on HAART	1	0	M	MSW
52.	26	.9691992	Still on HAART	2	0	M	MSW
53.	27	3.364819	Still on HAART	1	0	M	MSM
54.	27	3.364819	Still on HAART	2	0	M	MSM
55.	28	1.21013	Still on HAART	1	0	M	MSM
56.	28	1.21013	Still on HAART	2	0	M	MSM
57.	29	.3641342	TI	1	1	M	MSM
58.	29	.3641342	TI	2	0	M	MSM
59.	30	.936345	TI	1	1	M	MSM
60.	30	.936345	TI	2	0	M	MSM

Finally we declare our data as survival data and we perform the analysis.

```
stset time, f(fail)
```

b.1) The "stratified" version of the analysis can be done with the command:

```
xi: stcox i.gender*i.ft i.expo*i.ft, strata(ft)
```

The model fitted with the above command is:

$$\lambda_{\text{ft}}(t;\text{gender},\text{expo}) = \lambda_{\text{ft},0}(t)\exp\{(\beta_1 + \beta_2 I(\text{ft=2}))I(\text{gender=1}) + (\beta_3 + \beta_4 I(\text{ft=2}))I(\text{expo=1}) + (\beta_5 + \beta_6 I(\text{ft=2}))I(\text{expo=2}) + (\beta_7 + \beta_8 I(\text{ft=2}))I(\text{expo=3})\}$$

We can use robust estimates for the parameters' variances by adding the `robust` and `cluster` options:

```
xi: stcox i.gender*i.ft i.expo*i.ft, strata(ft)
robust cluster(id)
```

b.2) We can perform the "unstratified" version of the analysis by omitting the `strata` option:

```
xi: stcox i.gender*i.ft i.expo*i.ft
```

The model fitted with the above command is:

$$\lambda_{\text{ft}}(t;\text{gender},\text{expo}) = \lambda_{0}(t)\exp\{(\beta_1 + \beta_2 I(\text{ft=2}))I(\text{gender=1}) + (\beta_3 + \beta_4 I(\text{ft=2}))I(\text{expo=1}) + (\beta_5 + \beta_6 I(\text{ft=2}))I(\text{expo=2}) + (\beta_7 + \beta_8 I(\text{ft=2}))I(\text{expo=3}) + \beta_9 I(\text{ft=2})\}$$

Robust variance estimates can be used as previously by adding `robust` and `cluster` options:

```
xi: stcox i.gender*i.ft i.expo*i.ft, robust
cluster(id)
```

It should be noted that Abdel Babiker has written an ado program that performs the data augmentation method (`crcox.ado`).

In summary, we can fit a model for a single cause-specific hazard function by treating all observations that have failed from other causes as censored and fitting on the resulting dataset the usual Cox proportional hazards model. Additionally, we can fit a model for cause-specific hazard functions for all competing events simultaneously using the data augmentation method. In this case two alternative ways exist: **a)** to fit a Cox proportional hazards model stratified by failure type or **b)** to include the failure type as a covariate into the model. In both cases interactions of each covariate with failure type are also included in the model.

2.1.2 Implementation of the methods using R

To display the implementation of the methods in R the dataset described in Appendix 1 will also be used. Please note that when a dataset is imported from another statistical package caution is needed in the format variables (and particular factor variables) are stored. When, for example, we imported the dataset saved in STATA format, the variables with labels (event, expo and gender) were stored as factors. When the categorical variables are stored in R as numeric variables, we should declare them as factors by using `as.factor(variable name)` [instead of `variable name`] when we fit the models presented below.

To download R please refer to Appendix 2. Package `foreign` contains functions needed to read (and write) a dataset stored in various formats (STATA, SPSS etc). We can load this package by using `library` function:

```
library(foreign)
```

To create a data frame containing our data, which are stored in STATA in the present example, we must use the `read.dta` function as follows:

```
data<-read.dta("path\\changeHAART.dta")
```

In order to apply survival analysis in R we must load the package `survival`:

```
library(survival)
```

a) The model for the cause-specific hazard of TI can be fitted, as mentioned earlier, by defining as censored the observations with a new HAART regimen and using the `coxph` function:

```
data$ti<-ifelse(data$event=="TI",1,0)
#If subject's first change was TI then
#variable ti takes value 1, otherwise it takes 0

coxph(Surv(time,ti)~ gender+expo, data = data)
```

In R the default method for handling ties is Efron's approximation. If one wants to use Breslow's method (although Efron's approximation is prefered) argument `method` has to be added as follows:

```
coxph(Surv(time,ti)~gender+expo, data = data,
method="breslow")
```

b) The application of the data augmentation method needs the generation of the expanded dataset containing a failure type identifier (`ft`) for each record.

```
data1<-data.frame(data,ft=1)
#ft=1 corresponds to TI

data2<-data.frame(data,ft=2)
#ft=2 corresponds to new HAART regimen initiation

expanded<-rbind(data1,data2)
#Function rbind take a sequence of vector, matrix
#or data frames arguments and combine them by
#rows. Function cbind combines such a sequence by
#columns.
```

We additionally generate an indicator variable taking the value 1 in records corresponding to the actual failure of each subject. Of course censored subjects will have 0 in all their records.

```
expanded$event2<-
1*ifelse(expanded$event=="TI",1,0)+
2*ifelse(expanded$event=="new HAART",1,0)

expanded$fail<-
ifelse(expanded$event2==expanded$ft,1,0)
```

If interested, one could do the data management by using available on the internet functions. These can be found at `http://www.msbi.nl/multistate`.

b.1) The stratified version of the model can be fitted by the following command:

```
coxph(Surv(time,fail)~(gender+expo)*I(ft==2)+
strata(ft), data= expanded)
# I(.) is the indicator function
```

b.2) The unstratified version of the analysis can be performed, as mentioned before, by considering failure type (ft) as a covariate in the model (which just needs to omit the `strata` option):

```
coxph(Surv(time,fail)~(gender+expo)*I(ft==2),
data= expanded)
```

2.2 Cumulative incidence

Cumulative incidence for a particular cause of failure is the probability of experiencing this cause of failure until time t, in the presence of all the other possible causes. It should be noted that cumulative incidence for cause j is a subdistribution since, being defined as:

$$\lim_{t\to\infty} F_j(t) = \Pr(C = j)$$

It asymptotes at the probability that cause j is the first event. It is therefore not well defined statistically and modeling is more complex. A popular model for the cumulative incidence is the proportional hazards model for the subdistribution of a competing risk (Fine & Gray, 1999). This method makes use of the hazard of subdistribution which is a function of the cumulative incidence for the corresponding cause of failure and can be defined as:

$$\lambda_j^{sub}(t;x)=\lim_{h\to 0}\frac{1}{h}\Pr\{t\le T<t+h, C=j \mid T\ge t\cup(T\le t\cap C\ne j), x\}=$$
$$=\frac{\{dF_j(t;x)/dt\}}{\{1-F_j(t;x)\}}=-\frac{d\log\{1-F_j(t;x)\}}{dt}$$

The risk set of the above hazard is not natural since it includes at time t not only subjects who have not failed yet, but also subjects who have failed from other causes before t, who are not really at risk at that time. A semiparametric proportional hazards model is assumed for hazard of subdistribution and the cumulative incidence function has the form:

$$F_j(t;x)=1-\exp\left\{-\int_0^t \lambda_j^{sub}(u;x)\,du\right\}$$
$$\lambda_j^{sub}(t;x)=\lambda_{j0}^{sub}(t)\exp\{\beta_j' x\}$$

Estimation of the model parameters depends on 3 different scenarios two of which are not frequent in practice.

The first scenario concerns data from studies where all participants have "failed" from some cause, i.e. censoring is absent. In this case estimation for cause j can be achieved by defining observations with failures from other causes as censored and replacing observed corresponding times with a common value which is greater than the maximum failure time among all subjects that have experienced other causes than the cause of interest. In the resulting dataset, we can apply the Cox regression in any statistical package.

The second scenario involves data where censoring is only due to administrative termination of the study. The characteristic of such studies is that the potential censoring time is known even for subjects who have failed. The analysis requires to define observations with failures from other causes than j as censored and replace the corresponding times with the potential censoring time for each subject. Cox proportional hazards model can then be fitted in the resulting dataset.

Finally, the third scenario deals with data "suffering" from the usual random right censoring. Estimation here can be accomplished by the use of the Inverse Probability of Censoring Weighting technique (IPCW). Briefly, we define $r_i(t)=I(C_i\ge T_i\wedge t)$[1] taking 1 if it is known that subject i has not been censored or failed until t. This quantity takes 0 if status of subject i (i.e. has failed or not) is unknown at time t (i.e.

[1] $i\wedge j=\min(i,j)$

censoring has happened before both T_i and t). Based on this quantity and the estimated (using Kaplan-Meier estimator) survival distribution of the censoring random variable $\hat{G}(t) = \Pr(C > t)$ we define and associate in time t the time dependent weight $w_i(t) = r_i(t)\hat{G}(t)/\hat{G}(X_i \wedge t)$ where X_i is the minimum of T_i (failure time) and C_i. The weight $w_i(t)$ is equal to 1 if subject i has not failed neither has been censored until time t. If subject i has failed from another cause than j before t at time T_i, then the $w_i(t)$ is equal to $\hat{G}(t)/\hat{G}(T_i)$. It should be noted that the distribution of the censoring random variable can be related to a discrete covariate and this can be taken into account by estimating non-parametrically the weights separately for each group of the covariate. It is also possible to account for continuous covariates by assuming a proportional hazards model for the conditional distribution of C given these covariates. The method can be applied in R with the use of `cmprsk` package and the contained function `crr`. This function can estimate separately the weights within the levels of a categorical covariate, when this is necessary, although does not support continuous covariates as described above. Often, a combination of the second and third scenario holds.

2.2.1 Implementation of the methods using STATA

a) To illustrate the analysis under the first scenario we are going to pretend that censored individuals have changed their therapy to a new HAART regimen. Firstly, we define the failure indicator taking value 1 for individuals with TI and 0 otherwise, and then we replace the survival times in the individuals with new HAART initiation:

```
use changeHAART, clear

gen fail=(event==1)

qui sum time if event==1 /* we summarize failure
time for subjects with TI in order to get the
maximum time */

local m=r(max) /* we define local m to contain
the maximum failure time among subjects with TI
*/

replace time=`m'+1 if event!=1 /* we replace
failure times for individuals with other events
```

```
with a value higher than the maximum failure time
from the event of interest (i.e. TI) */
```

Now we can fit the Cox proportional hazards model:

```
stset time, f(fail)

xi: stcox i.gender i.expo, nohr /* reports the
estimated coefficients */
```

b) The second scenario is in principle applicable to our data because we know the date of last assessment for ART status for all patients, which coincides with the administrative censoring date. Therefore, we know the total length of the follow up even for individuals who change their first HAART scheme to a new HAART regimen (competing event). The basic assumption here is that the censoring distribution is not related to ART status. This, in our case, is a reasonable assumption. An exception could be censoring due to death (absorbing state) if time to death is associated with ART status. However, in our data, death is a rare event and even if independence does not hold for these subjects, the induced bias would be of limited size. So we first generate the failure indicator and replace failure times for individuals with new HAART initiation with their total follow up times:

```
use changeHAART, clear

gen fail=(event==1)

gen fup=(lastart-haartda)/365.25 /* Years from
first HAART initiation to the last assessment for
ART status */

replace time=fup if event==2 /* we replace event
times for subjects with initiation of a new HAART
regimen with the corresponding potential
censoring time */
```

Finally we can fit again the Cox proportional hazards model as usual:

```
stset time, f(fail)

xi: stcox i.gender i.expo, nohr /* reports the
estimated coefficients */
```

To our knowledge, the method of the third scenario has not yet been implemented in STATA.

2.2.2 Implementation of the methods using R

a) As in the corresponding STATA example to illustrate the first scenario we are going to assume that censored individuals switched to a new HAART regimen. First we generate the failure indicator taking the value 1 in individuals with TI and 0 otherwise and we replace times in the individuals with new HAART initiation:

```
data<-read.dta("path\\changeHAART.dta")

m<-max(data[data$event==”TI”,]$time)
or
m<-max(data[data$event==”TI”,”time”])
#Maximum time to TI

data$fail<-ifelse(data$event==”TI”,1,0)

data$time2<-
I(data$event!=1)*(m+1)+I(data$event==1)*data$time
#we replace failure times for individuals with
#other events with a value higher than the
#maximum failure time from the event of interest
#(i.e. TI)
```

Using data frame `data` we can fit Cox proportional hazards model:

```
coxph(Surv(time2,fail)~ gender+expo,data=data)
```

b) To apply the second scenario we must first calculate the total follow up time and then replace times in individuals started a new HAART regimen with the total follow up time:

```
data<-read.dta("path\\changeHAART.dta")

data$fup<-(data$lastart-data$haartda)/365.25
#Years from first HAART initiation to date of
#last assessment for ART
```

```
data$fail<-ifelse(data$event=="TI",1,0)

data$time2<-I(data$event=="new HAART")*data$fup+
I(data$event!="new HAART")*data$time
#we replace event times for subjects with
#initiation of a new HAART regimen with the
#corresponding potential censoring time
```

We can then fit the Cox proportional hazards model:

```
coxph(Surv(time2,fail)~ gender+expo,data=data)
```

c) Applying the analysis under the third scenario, in our example implies that we do not make use of the knowledge of the potential censoring time for individuals initiated a new HAART regimen, requires loading the `cmprsk` package:

```
library(cmprsk)
```

This package includes various useful functions to be used in a competing risk analysis. The included function `crr` fits the proportional hazards model under usual right censoring (weighted estimating equations). The syntax of this function is:

```
crr(ftime, fstatus, censor, cov1, cov2, tf,
cengroup, failcode=1, cencode=0, subset,
na.action=na.omit, gtol=1e-06, maxiter=10, init)
```

`ftime` is a vector containing failure/cencoring times (see `time` in Appendix 1) and `fstatus` is another vector containing failure types (or censoring) (see `event` in Appendix 1). `cov1` is a matrix (number of observations x number of covariates) of fixed covariates and `cov2` is another matrix of covariates to be multiplied with functions of time defined in `tf`. For example if we wanted to fit a quadratic in time model for a covariate z (i.e. $\beta_1 z+\beta_2 zt+\beta_3 zt^2$) we should specify `cov1<-z, cov2<-cbind(z,z), ft=function(uft) cbind(uft, uft*uft)` where `uft` is the vector of unique failure times with failure from cause of interest. If the censoring distribution is different between different groups then we must specify `cengroup` which is a vector indicating these groups. `failcode` and `cencode` indicate the code/value of the failure type we are interested for and the code/value for censoring. `subset` is a logical vector restricting the analysis in a subset

of subjects. In `na.action` we optionally specify the action to take for any cases missing any of ftime, fstatus, cov1, cov2, cengroup, or subset. Finally, in `maxiter` and `init` we optionally define the maximum number of iterations in Newton algorithm and the initial values of regression parameters.

In our example we must first construct the vectors of failure times and types and the matrix of covariates. In our application we will generate the dummies for exposure group (see Appendix 1) in order to include them in the matrix of the independent variables (`cov1`) that will be entered in the model.

```
data<-read.dta("path\\changeHAART.dta")

data$event2<-1*ifelse(data$event=="TI",1,0)+
2*ifelse(data$event=="new HAART",1,0)

#Generation of dummies
g1<-ifelse(data$gender=="F",1,0)

e2<-ifelse(data$expo=="IDU",1,0)
e3<-ifelse(data$expo=="MSW",1,0)
e4<-ifelse(data$expo=="HAE/OTH-UNK",1,0)

cov<-cbind(g1, e2, e3, e4)

#Alternative way of generating the dummies for
#expo
e<-factor(data$expo)
e<-model.matrix(~e-1)
e<-e[,c(2:4)] # first column (expo==0-MSM) is the
#reference category

cov<-cbind(g1, e)
```

Then we can fit the model using the `crr` function:

```
crr(data$time, data$event2, cov, cencode=0,
failcode=1)
```

The output reports estimated coefficients, SEs and p-values:

```
convergence:  TRUE
coefficients:
```

```
[1]  0.40540  0.58030 -0.02021  0.06800
standard errors:
[1] 0.1733 0.1483 0.2118 0.2472
two-sided p-values:
[1] 1.9e-02 9.1e-05 9.2e-01 7.8e-01
```

3. Conclusions

As conclusions, a few points need to be highlighted.

a. There is nothing wrong with modeling cause-specific hazard functions by fitting the usual Cox proportional hazards model after defining as censored failures from other causes than the one we are interested in. This quantity, as mentioned in section 1.3, is estimable from the observable (competing risks) data and does not depend on further untestable assumptions about the joint and marginal distributions of the latent failure times for the possible events (such as independence). However, it does not model the marginal hazard, which is the one that is modeled in a standard survival analysis.

b. On the other hand, the other well known tool of classical survival analysis, the Kaplan-Meier estimator, cannot be used in the competing risks setting. It can be proven that the quantity 1-$\hat{S}'_j(t)$ (where $\hat{S}'_j(t)$ is the Kaplan-Meier estimate we obtain if we define as censored the observations with failures from other causes than j) which has been used frequently in the past, overestimates the cumulative incidence and it can result in cumulative incidences that add (sum over all competing events) to a number higher than 1. Non-parametric estimation of cumulative incidence can be implemented in R with function `cuminc` which is contained in package `cmprsk`, and in STATA with `stcompet.ado` which can be found on the internet or by using the `findit` command.

c. The analysis of the two basic, estimable from the data, quantities (i.e. cause-specific hazard function and cumulative incidence) can give different results. This is due to the fact that the effect of a covariate on cause-specific hazard function for a particular cause can be different from its effect on the cumulative incidence of the corresponding cause (see sections 1.1.2 & 2). Thus, one must choose the desired quantity for the data, in terms of interpretation and importance from a biomedical perspective.

Appendix 1. Description of the dataset

Briefly, we are interested in time to TI or shift to a new HAART regimen after having been in a stable HAART for at least 90 days. The variable we used were: `time` (years) from 1st stable HAART (`haartda`) to the 1st major change in ART (i.e. TI or new HAART regimen initiation) or until date for last assessment of ART status (`lastart`) for subjects without such a change. Type of change in ART status or study termination without a change was included in `event` whereas the variable `expo` provides the possible source of infection (or exposure group).

```
-------------------------------------------------------------------------------
time                                                    <<Failure>> time
-------------------------------------------------------------------------------

                  type:  numeric (float)

                 range:  [.24640657,6.7515402]        units:  1.000e-08
          unique values:  916                      missing .:  0/1551

                  mean:   2.04653
             std. dev:    1.428

            percentiles:        10%       25%       50%       75%       90%
                             .501027   .895277   1.66188   2.91581   4.17522

-------------------------------------------------------------------------------
event                                                   1st major change in HAART
-------------------------------------------------------------------------------

                  type:  numeric (float)
                 label:  evlbl

                 range:  [0,2]                        units:  1
          unique values:  3                        missing .:  0/1551

             tabulation:  Freq.   Numeric  Label
                            969         0  Still HAART
                            299         1  TI
                            283         2  New HAART

-------------------------------------------------------------------------------
gender                                                              (unlabeled)
-------------------------------------------------------------------------------

                  type:  numeric (float)
                 label:  sexlbl

                 range:  [0,1]                        units:  1
          unique values:  2                        missing .:  0/1551

             tabulation:  Freq.   Numeric  Label
                           1280         0  M
                            271         1  F
```

```
-------------------------------------------------------------------------------
expo                                                  Possible source of infection
-------------------------------------------------------------------------------

                 type:  numeric (float)
                label:  expolbl

                range:  [0,3]                        units:  1
        unique values:  4                          missing .:  0/1551

           tabulation:  Freq.   Numeric  Label
                          869         0  Men having sex with men (MSM)
                          349         1  Inductive drug users (IDU)
                          244         2  Men having sex with women (MSW)
                           89         3  Hemophiliacs, other-unknown




-------------------------------------------------------------------------------
haartda                                                 Date of 1st stable HAART
-------------------------------------------------------------------------------

                 type:  numeric daily date (int)

                range:  [12774,15768]                units:  1
       or equivalently:  [22dec1994,04mar2003]       units:  days
        unique values:  889                        missing .:  0/1551

                 mean:   14241.7 = 28dec1998 (+ 16 hours)
              std. dev:   563.283

          percentiles:        10%       25%       50%       75%       90%
                             13588     13808     14167     14626     15088
                          15mar1997 21oct1997 15oct1998 17jan2000 23apr2001

-------------------------------------------------------------------------------
lastart                                      Date of last assessment of ART status
-------------------------------------------------------------------------------

                 type:  numeric daily date (int)

                range:  [13319,15929]                units:  1
       or equivalently:  [19jun1996,12aug2003]       units:  days
        unique values:  762                        missing .:  0/1551

                 mean:   15282.5 = 03nov2001 (+ 12 hours)
              std. dev:   541.886

          percentiles:        10%       25%       50%       75%       90%
                             14440     14956     15474     15704     15823
                          15jul1999    12dec2000    14may2002    30dec2002    28apr2003
```

Appendix 2. R software and package downloads

R statistical software as well as `cmprsk` package can be downloaded from http://www.r-project.org/. After choosing "CRAN" (Figure 1a) we select a mirror (Figure 1b) and then the operating system of the computer in which R is going to be installed (Figure 1c). There we can download the software (Figures 1d-e). Packages, including `cmprsk`, can be downloaded and installed via "intall package(s)" from "package" menu.

Figure 1a.

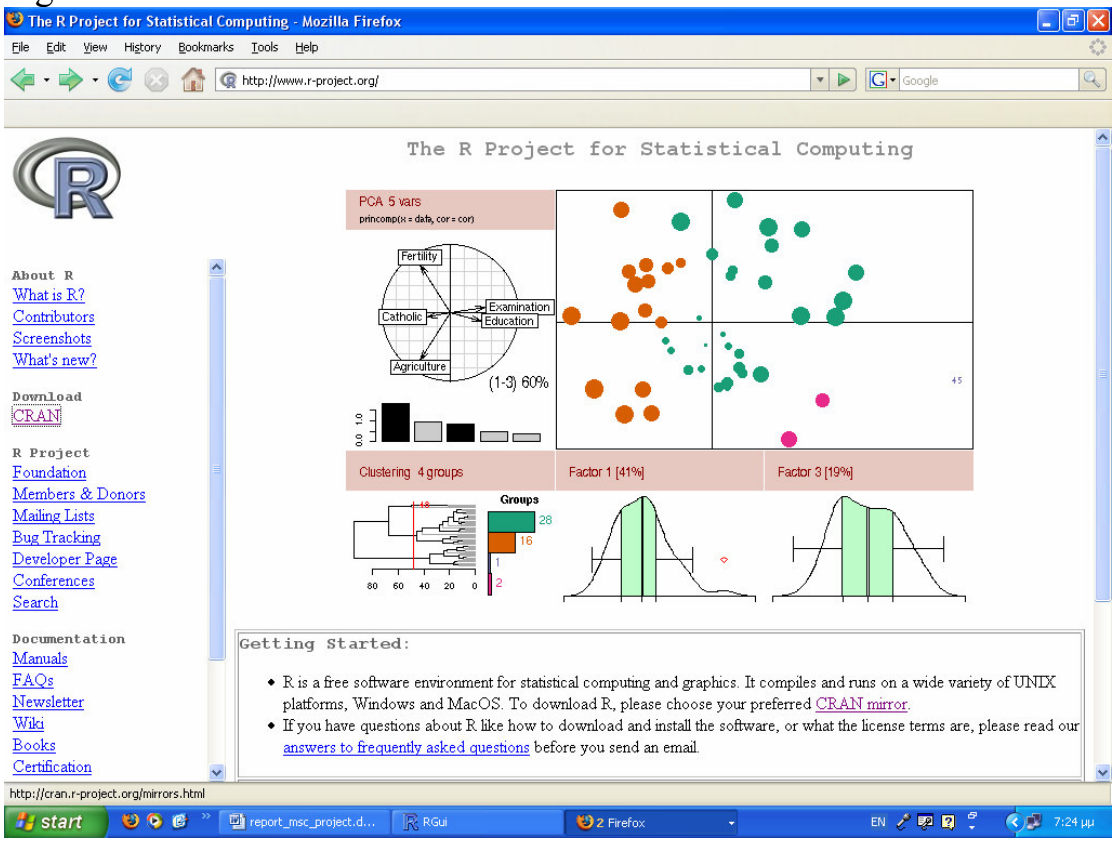


Figure 1b.

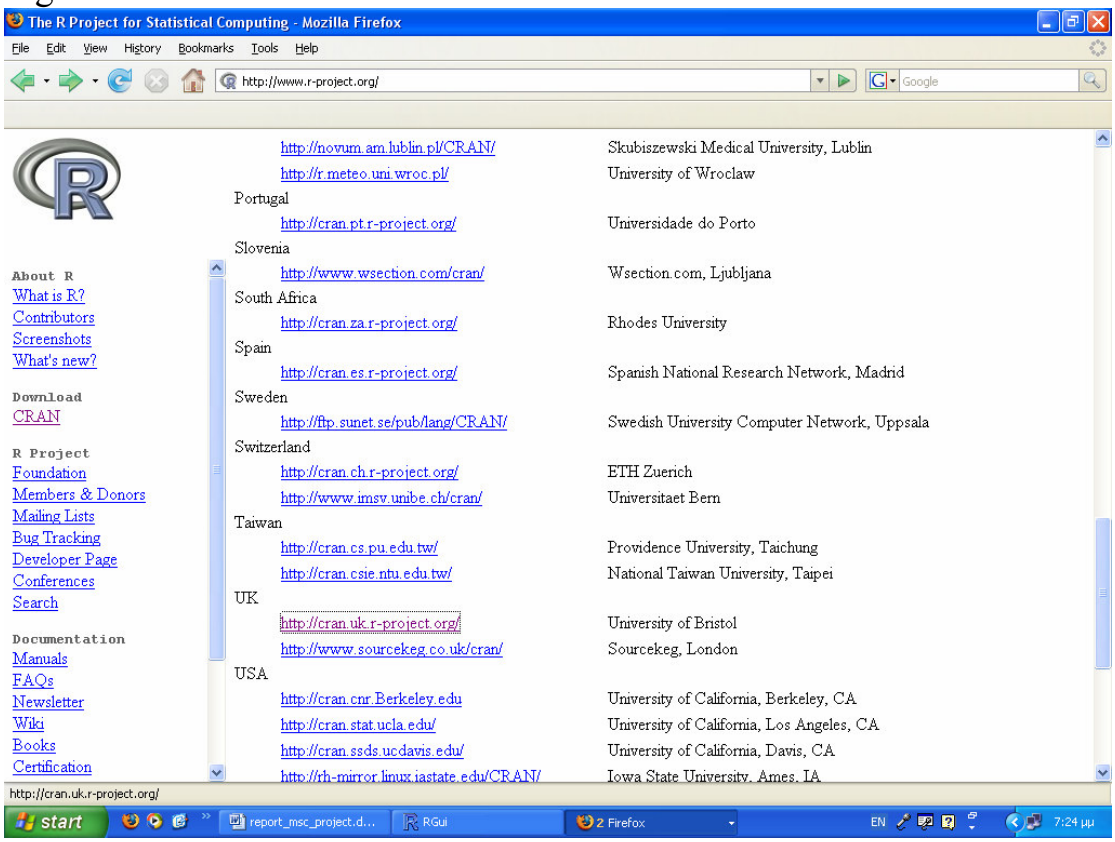


Figure 1c.

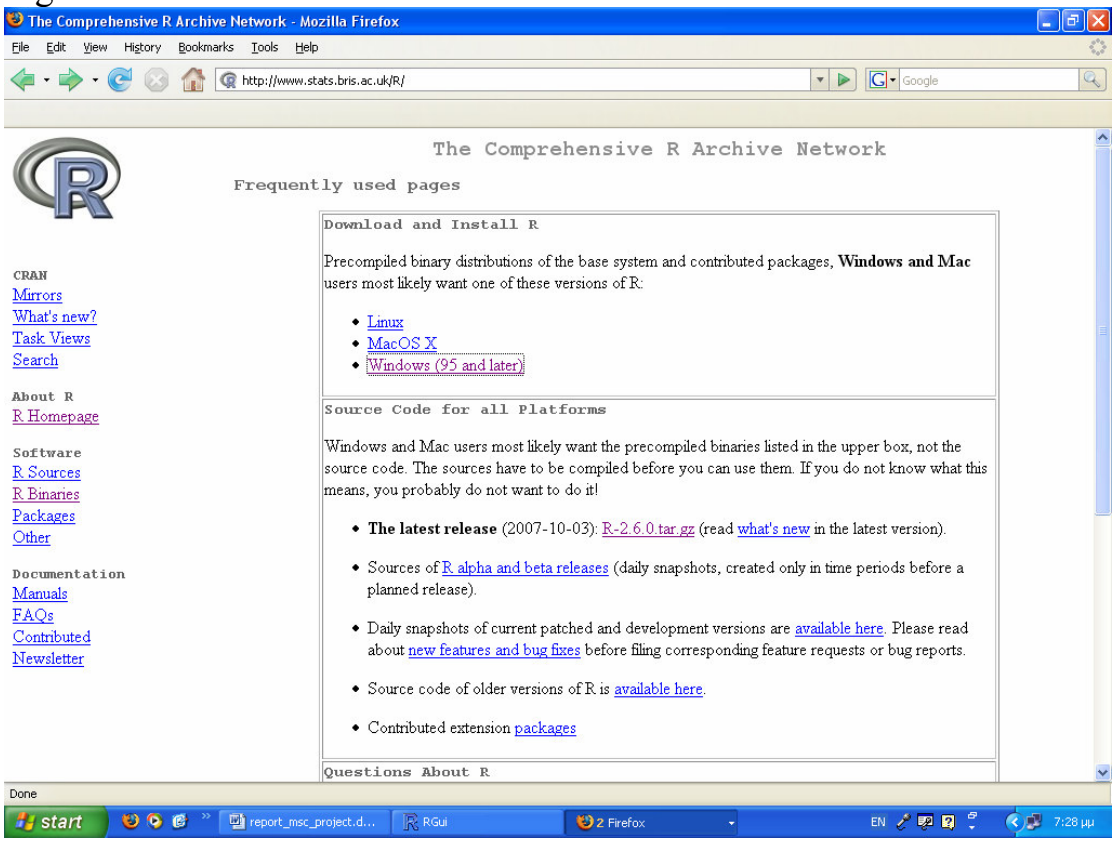


Figure 1d.

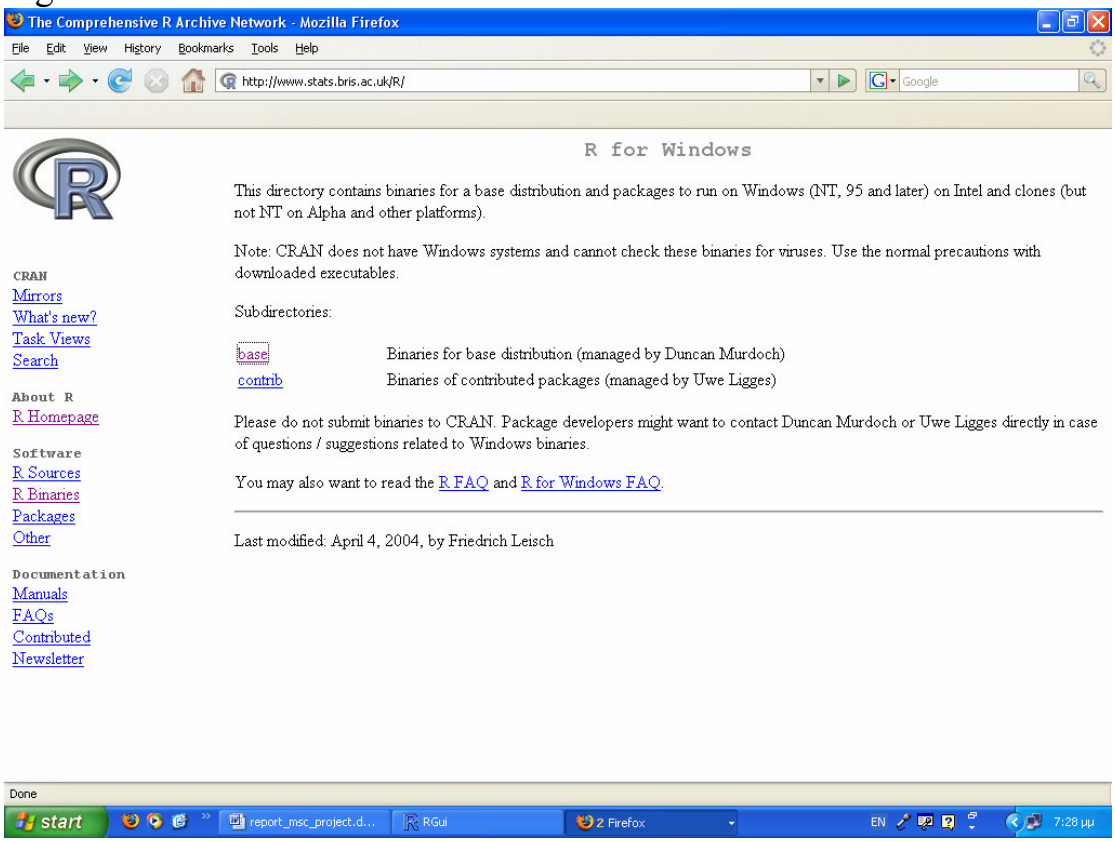
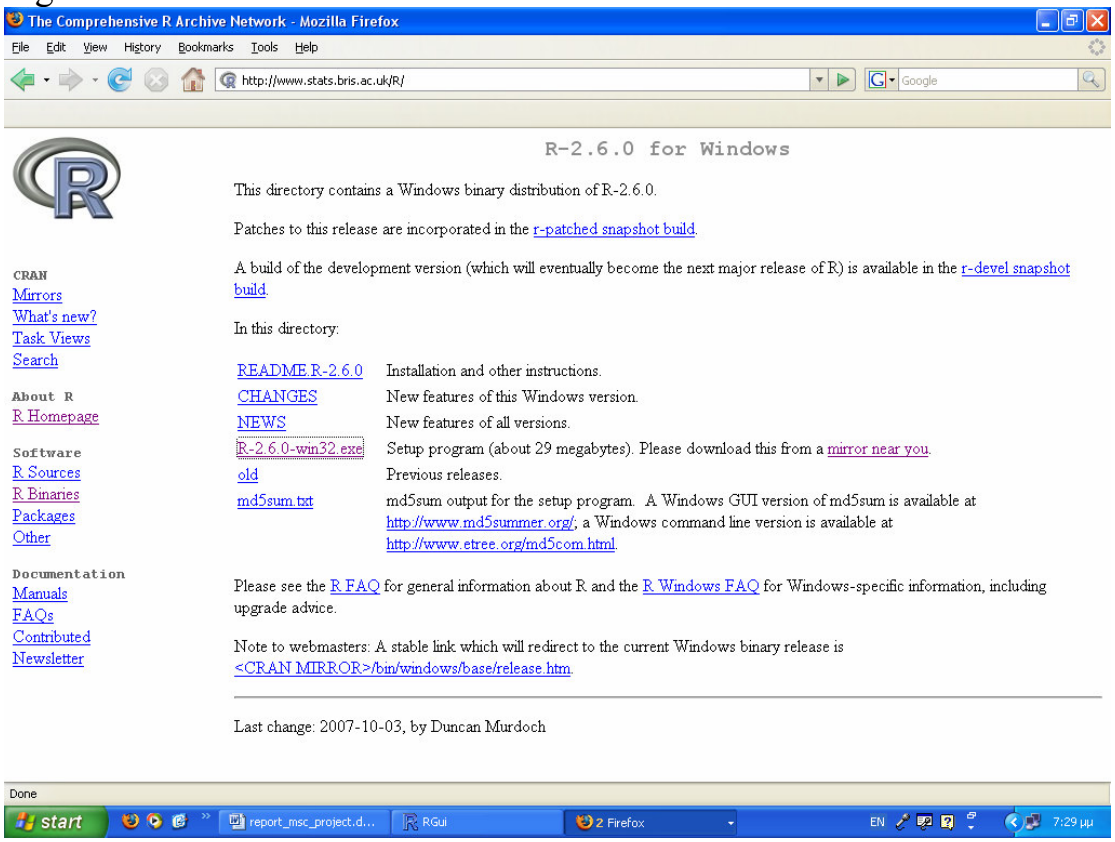


Figure 1e.



Appendix 3. Comparison of the methods

As part of Mr. Bakoyannis' M.Sc. in Biostatistics summer project, methods described in this report were applied to analyze CASCADE data presented in Appendix 1. As already mentioned, potential censoring time, which coincides with date of last assessment of ART status, was known for all individuals, even those with a failure (i.e. major ART change). As a result we replaced, as described above, the event times of those who had started new HAART with total length of their follow-up and fitted a Cox proportional hazards model with failure indicator taking 1s in observations with TI. A backward selection process was used for model building and the results are presented in Table 1. This table also includes results from the fitted model for the cause specific hazard function of TI and of initiation of a new HAART regimen.

Table 1. Results [$\hat{\beta}\ (p-value)$] of subdistribution hazard model for TI and cause-specific hazard models for TI and switching to a new HAART regimen (NH).

	$\lambda_{TI}^{sub}(t)$	$\lambda_{TI}(t)$	$\lambda_{NH}(t)$
Gender			
Female/Male	0.475 (0.006)	0.522 (0.003)	0.153 (0.472)
Source of infection			
IDU/MSM	0.503 (0.001)	0.397 (0.01)	-0.373 (0.032)
MSW/MSM	-0.068 (0.747)	-0.104 (0.626)	-0.109 (0.611)
other/MSM	0.089 (0.725)	-0.085 (0.737)	-0.772 (0.006)
Virologic response to 1st HAART			
Initial/Sustained	0.191 (0.164)	0.304 (0.026)	1.011 (<0.001)
None/Sustained	0.584 (<0.001)	1.12 (<0.001)	2.418 (<0.001)
1st HAART baseline HIV-RNA			
per 1 log10 copy/ml	0.16 (0.007)	0.166 (0.005)	0.012 (0.847)
Year of 1st HAART initiation			
98-99/<98	0.324 (0.02)	0.292 (0.034)	-0.057 (0.658)
00-03/<98	0.469 (0.01)	0.44 (0.014)	-0.52 (0.018)

As we can see in Table 1, results from the two methods are in general similar (minimal or small differences in the corresponding effect estimates). The only exception is the estimate for "virologic response". There is no evidence (p=0.164) of a difference between initial and sustained responders to 1st HAART regimen for cumulative incidence,

but a significant difference exist (p=0.026) for cause-specific hazard for TI. Moreover, the difference between no and sustained responders is greater for cause-specific hazard of TI ($\hat{\beta}$=1.12) than for hazard of cumulative incidence ($\hat{\beta}$=0.584). This is due to the fact that virologic response seems to affect strongly also the cause-specific hazard for new HAART initiation and this effect is stronger than the corresponding effect on cause-specific hazard for TI. As we have already noted, the effect of a covariate on the cumulative incidence for one cause does not depend only on the effect of the covariate on the cause-specific hazard for the corresponding cause, but also on the respective effect on the cause-specific hazard functions for all the other causes as well as on the baseline hazard functions.

To better illustrate this effect we did the following exercise: we estimated the cumulative incidence of TI for males and females based on the results from the cause-specific hazard models (for TI and new HAART regimen initiation) presented in Table 1 but varying the gender effect on cause-specific hazard for initiating a new HAART. More specifically, the effect estimate was ranged from 0 (no effect of gender on cause-specific hazard of new HAART) to 1.75 (strong effect with females having higher instantaneous failure rate than males to switch to a new HAART regimen). Five different scenarios were examined determined by the gender effect size ($\hat{\beta}_{NH}^{gender}$ = 0 or 0.75 or 1.25 or 1.5 or 1.75). The results are graphically presented in Figures 2a-2e. It should be noted that the effect estimates of all other covariates presented in Table 1 were kept fixed in all scenarios.

Figure 2a. Estimated cumulative incidence of TI for males and females ($\hat{\beta}_{NH}^{gender}$=0)

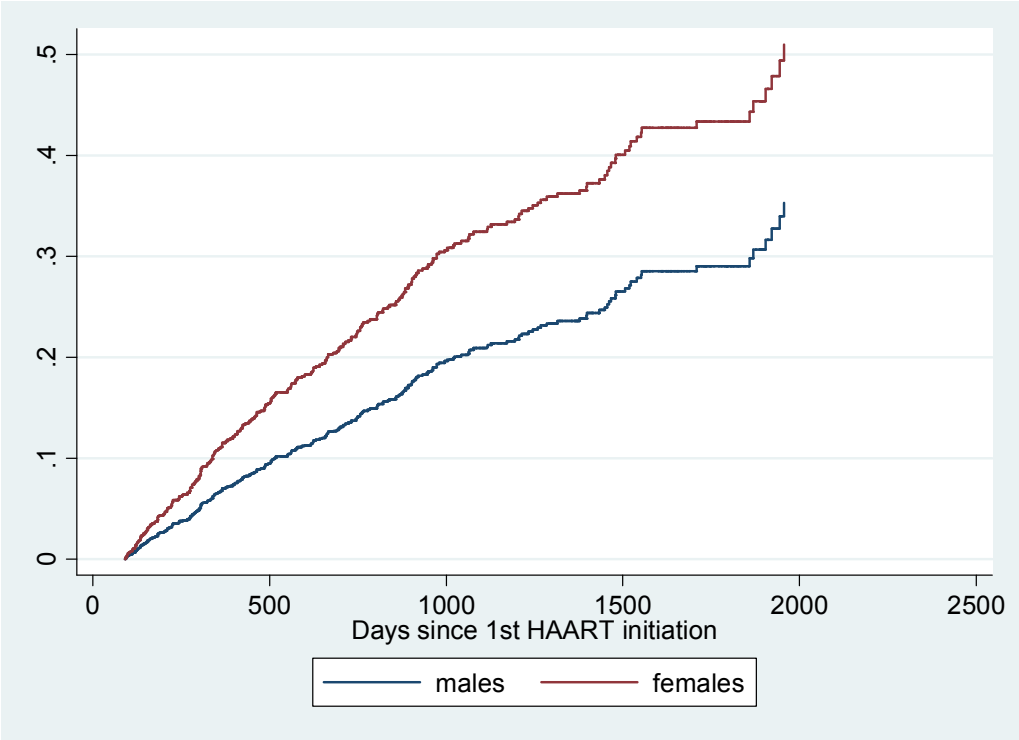


Figure 2b. Estimated cumulative incidence of TI for males and females ($\hat{\beta}_{NH}^{gender}$=0.75)

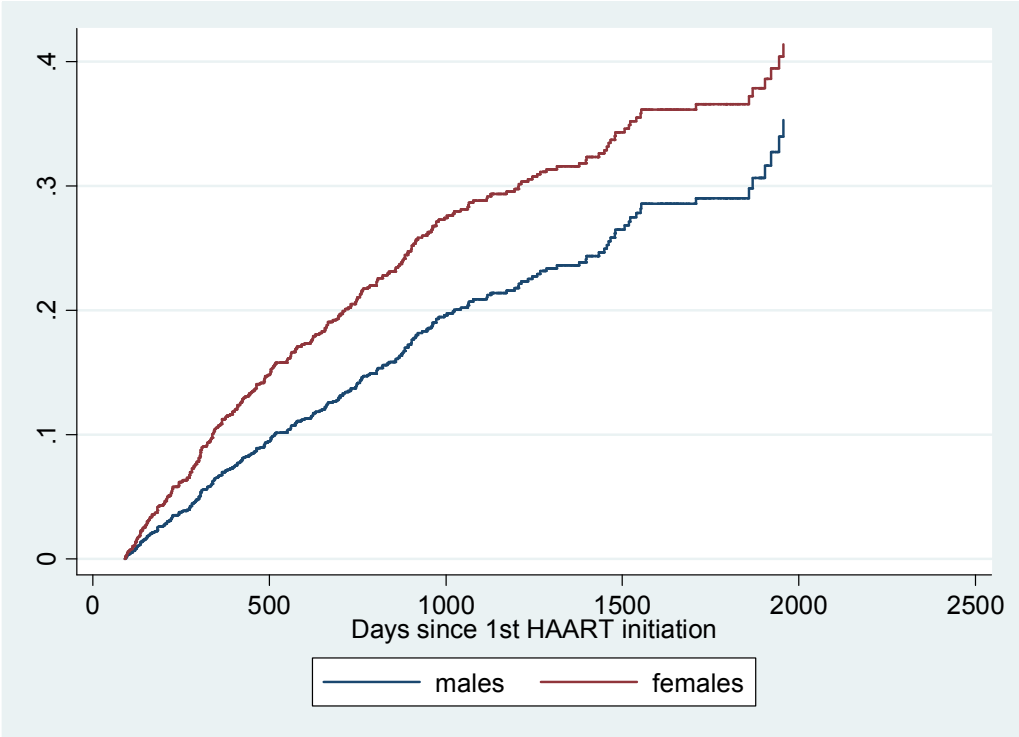


Figure 2c. Estimated cumulative incidence of TI for males and females ($\hat{\beta}_{NH}^{gender}$=1.25)

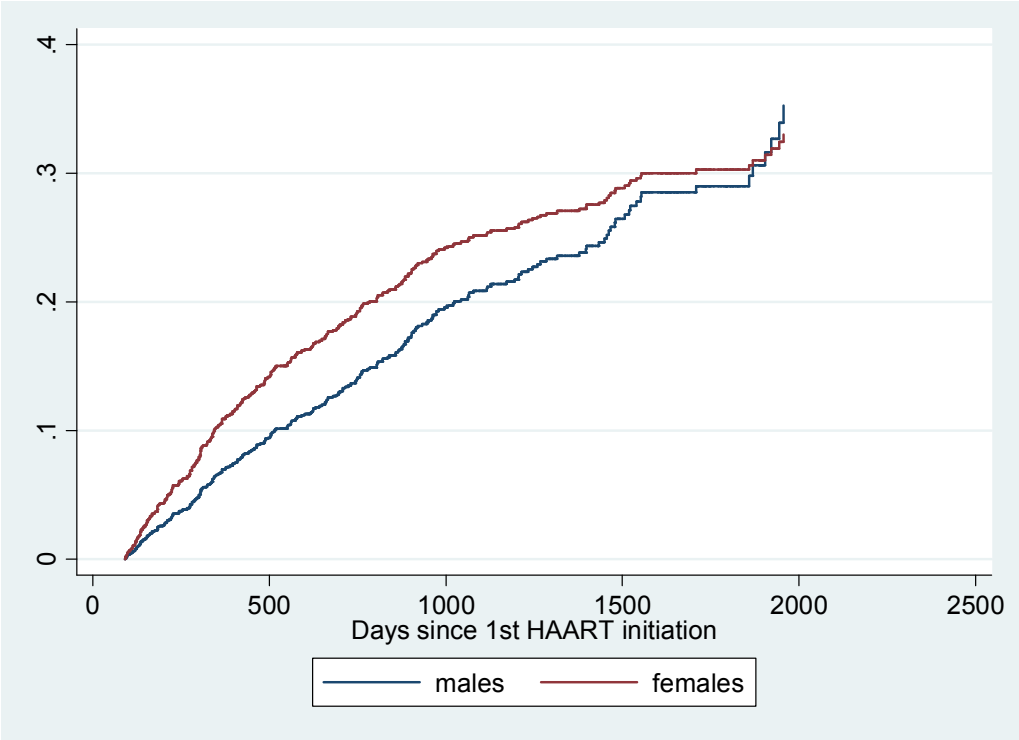


Figure 2d. Estimated cumulative incidence of TI for males and females ($\hat{\beta}_{NH}^{gender}$=1.5)

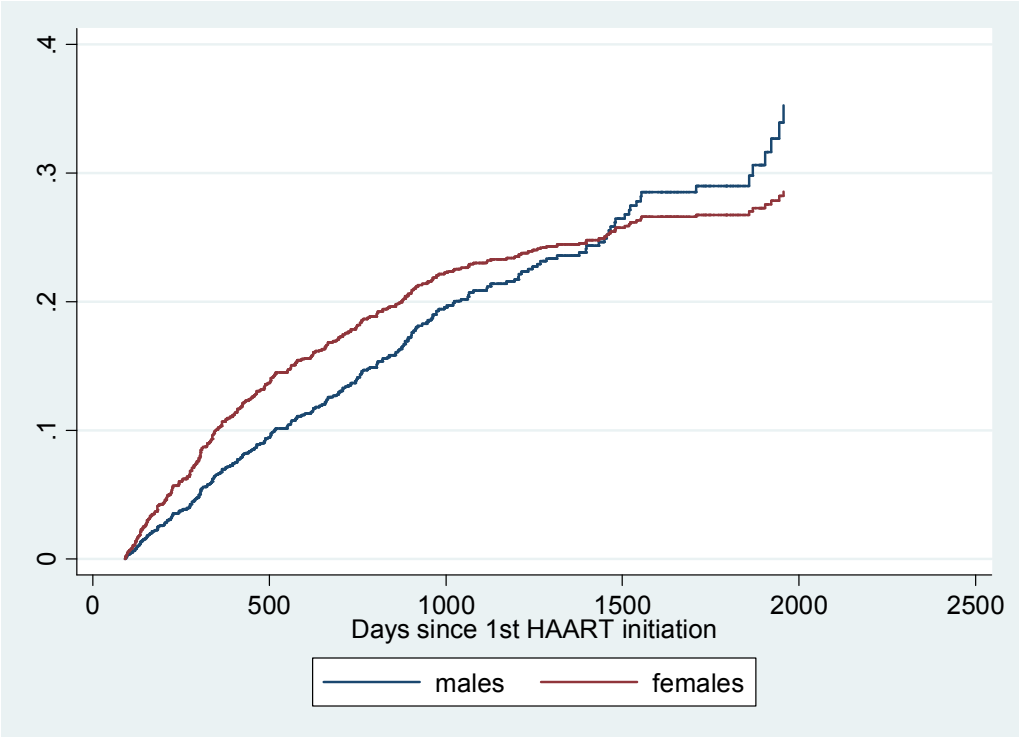
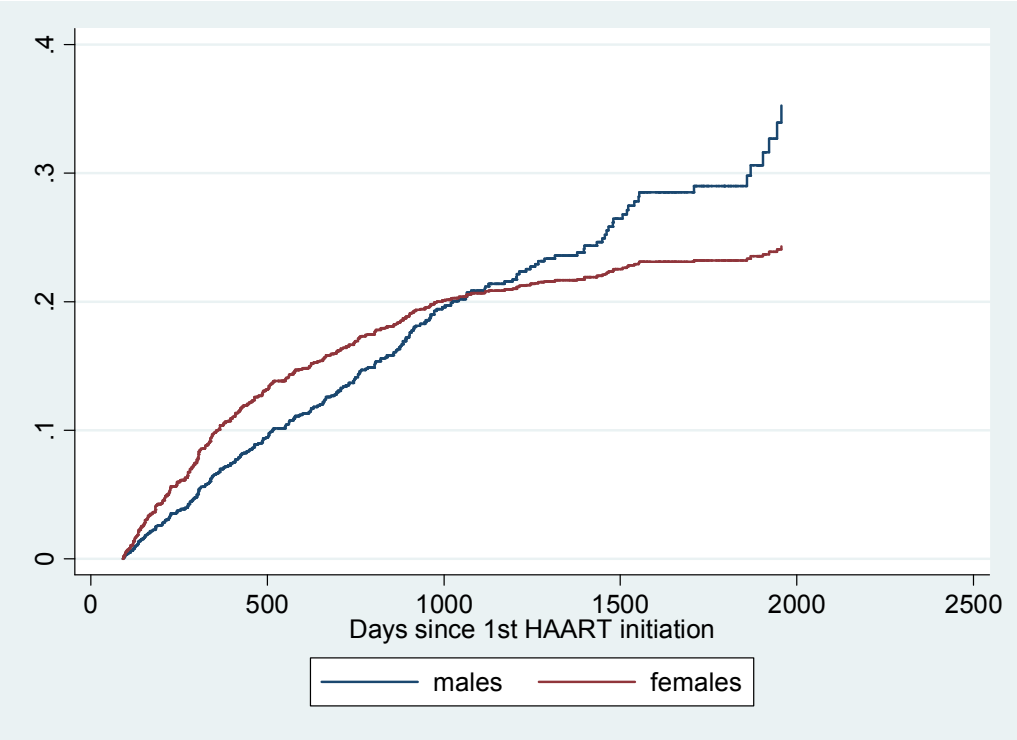


Figure 2e. Estimated cumulative incidence of TI for males and females ($\hat{\beta}_{NH}^{gender}$=1.75)



As it can be seen in Figures 2b-2e, as the gender effect on cause-specific hazard of new HAART initiation becomes stronger, that is women are more and more likely than men to initiate a new HAART regimen, the by gender difference in the cumulative incidence of TI becomes smaller over time and finally the two curves cross.

If we ignore the knowledge of the potential censoring time, we would fit the model for the subdistribution hazard of TI using the weighted estimation equation approach. In such a case, if the censoring distribution depends on some covariates, this has to be taken into account when estimating the weights used in this method. R function `crr` can take this fact into account (if the covariate or the covariates are categorical) by estimating the weights separately within groups with different censoring distributions. In our data censoring distribution seems to be affected by type (PI or non-PI based) and year of initiation of 1st HAART initiation[2]. A 6 level categorical covariate was generated by the combination of levels of type and year of 1st HAART and used in `cengroup` argument of `crr` function. In Tables 2a and 2b results from the three different methods for fitting the subdistribution hazard

[2] Please not that the purpose of our example is to illustrate the application of the methods rather than to draw appropriate inferences. Therefore results presented here should be interpreted with cautiously (especially those referring to PI vs non-PI comparisons).

model for TI are presented. The first one makes use of the knowledge of the potential censoring time, whereas the other two ignore this information and apply the weighted estimating equation approach. The second scenario ignores the dependence of the censoring distribution on the type and year of 1st HAART whereas the third takes this into account.

Table 2a. Results [$\hat{\beta}$ (% change from corresponding $\hat{\beta}$ derived from the "censoring complete" method)] from modeling the subdistribution hazard of TI applying different methods to deal with censoring.

	cc[1]	urc[2]	urc_dw[3]
Gender			
Females/Males	0.475	0.474 (-0.215)	0.480 (0.934)
Source of infection			
IDU/MSM	0.503	0.482 (-4.229)	0.481 (-4.443)
MSW/MSM	-0.068	-0.092 (35.092)	-0.078 (14.122)
other/MSM	0.089	0.084 (-5.314)	0.088 (-1.166)
Virologic response			
Initial/Sustained	0.191	0.187 (-2.162)	0.192 (0.512)
None/Sustained	0.584	0.555 (-5.005)	0.559 (-4.188)
1st HAART baseline HIV-RNA			
per 1 log10 copy/ml	0.160	0.173 (8.504)	0.170 (6.619)
Year of 1st HAART initiation			
98-99/<98	0.324	0.249 (-23.288)	0.310 (-4.454)
00-03/<98	0.469	0.321 (-31.582)	0.448 (-4.329)

[1] Censoring complete
[2] Usual right censoring
[3] Usual right censoring taking into account the differentiation of the censoring distribution

Table 2b. Standard errors for the estimated coefficients derived from modeling the subdistribution hazard of TI applying different methods to deal with censoring.

	cc[1]	urc[2]	urc_dw[3]
Gender			
Females/Males	0.171	0.177	0.177
Source of infection			
IDU/MSM	0.153	0.155	0.164
MSW/MSM	0.212	0.217	0.222
other/MSM	0.253	0.249	0.373
Virologic response			
Initial/Sustained	0.137	0.135	0.157
None/Sustained	0.152	0.154	0.256

1st HAART baseline HIV-RNA			
per 1 log10 copy/ml	0.059	0.061	0.063
Year of 1st HAART initiation			
98-99/<98	0.139	0.136	0.139
00-03/<98	0.181	0.175	0.179

[1] Censoring complete
[2] Usual right censoring
[3] Usual right censoring taking into account the differentiation of the censoring distribution

As shown in Table 2a taking into account the differentiation of censoring distribution in the estimation of the weights generally leads to estimates that are closer to those obtained by using the knowledge of the potential censoring times. This is particularly evident for the estimates of heterosexuals (% differences between the "censoring complete" and "usual right censoring" method with common and different censoring distribution: 35.1 % and 14.1 % respectively). On the other hand as expected since in the third approach more quantities need to be estimated (i.e. the different censoring distributions for different groups) the parameters' associated SEs were somewhat larger than the corresponding SEs from the other two approaches.

The conclusions from the above exercises are:

1) The results from cause-specific hazard and cumulative incidence models for a specific cause of failure can differ substantially. Which one method (i.e. modeling cause-specific hazard function or cumulative incidence) is more appropriate would depend on the biological question to be answered.
2) When the potential censoring time is not known and there is evident that the censoring distribution depends on one or more covariates, this information should be taken into account when estimating the weights for the third scenario. However, one should remember that there is a cost for doing that: the larger SEs, but this cost is well outweighed by the smaller bias of the estimates.

References

Crowder MJ (2001). *Classical Competing Risks.* Chapman & Hall/CRC.

Fine JP, Gray RJ (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* 94:496-509.

Gaynor J, Feuer E, Tan C, Wu D, Little C, Straus D, Clarkson B, Brennan M (1993). On the Use of Cause-Specific Failure and Conditional Failure Probabilities: Examples From Clinical Oncology Data. *Journal of the American Statistical Association* 88:400-409.

Kalbfleisch JD, Prentice RL (2002). *The Statistical Analysis of Failure Time Data* (2nd edn). Wiley: New York.

Latouche A, Beyersmann J, Fine JP (2007). Letter to the editor: Comments on 'Analysing and interpreting competing risk data'. *Statis Med* 2007; 26:3676–3680.

Lunn M, McNeil D (1995). Applying Cox regression to competing risks. *Biometrics* 51:524-532.

Pintilie M (2007). Analysing and interpreting competing risk data. *Statist Med* 26:1360-1367

Putter H, Fiocco M, Geskus RB (2007). Tutorial in biostatistics: Competing risks and multi-state models. *Statist. Med.* 26:2389-2430.

R Development Core Team (2006). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Touloumi G, Pantazis N, Antoniou A, Stirnadel HA, Walker SA, Porter K; CASCADE Collaboration (2006). Highly active antiretroviral therapy interruption: predictors and virological and immunologic consequences. *J Acquir Immune Defic Syndr* 42:554-61.

Tsiatis AA (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences U.S.A.* 72:20-22.