How to create a *Biograph* object?

Frans Willekens

March 2012

1. Introduction

The purpose of this note is to show how to create a *Biograph* object. The R code is included in the *Biograph* package (Version 2). First I consider two sets of hypothetical data. The first data set carries information on three subjects, the second on 22 subjects. Next, I consider the German Life History Survey data, which are used throughout the book. I show how to obtain a *Biograph* object from the data on 201 individuals included in the subsample used by Blossfeld and Rowher (2002) to illustrate techniques of event history modeling. Data from the Netherlands Family and Fertility Survey 1989 are considered next. The fifth example of how to prepare a *Biograph* object uses data from the Survey of Health, Ageing and Retirement in Europe (SHARE). The SHARE survey is modeled after the U.S. Health and Retirement Survey (HRS). The sixth example uses data from the National Family Health Survey of India, which is one of many Demographic and Health Surveys (DHS) organized in third-world countries and countries in transition. The final example uses medical data included in the *mstate* package for multistate modeling in R, developed by Putter and colleagues at Leiden University Medical Centre. The data 2279 leukemia patients who had a bone marrow transplant.

The Biograph object carries information on subjects, on transitions and on the observation period. The subject data consist of the date of birth and a set of attributes (covariates). Time-varying attributes are included by the age(s) at which the attributes change. In the GLHS data, being ever married is a time-varying attribute. The age at marriage is included in the data. Information on transitions includes the state sequence and the ages at transition. The ages are ordered chronologically, with the age at the first event displayed first, followed by the age at the second event, etc. The ordering is consistent with the state sequence, which is the sequence of states occupied by the subject from start to end of observation.

The *Biograph* object is created in five steps. The first is the specification of the state space and the transitions. Transitions that are not relevant for the study are excluded. The transitions that are included are feasible and relevant. The second step is the selection of the covariates. The observation window for each subject in the observation is specified in the third step. It requires the dates at start and end of observation. In the fourth step, the state sequence is determined and the ages at transition are recorded. In the fifth and final step, all data are stored in a data frame and two data attributes are attached. The first is the format of the transition dates and the second is the transition matrix, i.e. the matrix of possible and relevant transitions. That matrix gives also information on the state space.

2. Hypothetical data A

Consider three individuals, one male and two females. Two have medium levels of education and one completed higher education. The three individuals are born in 1986. The first person is born on 5th April 1986, the second on 8th August 1986 and the third on 28th November 1986. Assume that during an interview on 9st May 2012 life history data were collected on living arrangement. The following living arrangements are distinguished: living at the parental home (H), living alone (A), cohabiting (C) and married (M). The first person (female) started living independently in August 2004 at the age of 18. It was her first transition, i.e. she left the parental home to live independently. She started cohabitation in December 2011 and is still cohabiting at the time of interview. The second person (male) started living independently in September 2011. The third person (female) started living independently in August 2006 and married on 16th March 2012. If the month of transition is known, but not the date, it is assumed that the transition takes place on the 15th of that month. The information on the transitions is summarized below:

```
           A          C          M
1 2004-08-15 2011-12-15       <NA>
2 2011-09-15       <NA>       <NA>
3 2006-08-15       <NA> 2012-03-16
```

A row carries information on an individual. A column has the date of entry in a given state. Consider the first person. She starts living alone on 15th August 2004 and starts cohabitation on 15th December 2011.

The state space is the set of possible living arrangements. It is the set {H, A, C, M}. The covariates are sex and level of education. The observation window differs for each individual. It starts at birth and ends at interview. The data are shown in Table 1

Table 1. Data retained for Biograph object: steps 1 - 3

```
  ID      start        end sex   educ
1  1 1986-04-05 2019-05-09   F   High
2  2 1986-08-08 2019-05-09   M Medium
3  3 1986-11-28 2019-05-09   F Medium
```

The first column is the line number. The second column is the subject's identification number (ID). The third and fourth columns delineate the observation window. The dates are objects of class 'Date', which enables arithmetic and logical operations on the dates. The fifth and sixth columns show the covariates. The covariates are factors.

The fourth step in the preparation of a *Biograph* object results in state sequences and the transition dates. To determine the state sequence, the transition dates need to be ordered chronologically, i.e. the event that occurred first is listed first. The subsequent event is listed second, etc. The second event is not the same for everyone. In the data above, it is cohabitation for the first person and marriage for the third person. The function `Sequences.ind.0` orders the dates chronologically and derives the state sequence. The raw transition dates (shown above) are stored in a data frame with the dates as character variables. The function `as.Date` of Base R is used to convert the character dates in Julian dates. The function is evoked using the code:

```
f <- Sequences.ind.0(d=dd,namstates=namstates,absorb=NULL)
```

where dd is the data frame with the transition dates and namstates is the state space. The function produces an object with several components, but two are of particular importance. They are the state sequence (`f$path`) and the sorted transition dates (`f$d`). Table 2 shows the object produced by the function `Sequences.ind.0`. The components `f$d`, `f$path` and `f$ns` are included in the *Biograph object*.

Table 2. Object produced by the function `Sequences.ind.0`

```
$namstates
[1] "H" "A" "C" "M"

$d
      [,1]  [,2] [,3]
[1,] 12645 15323   NA
[2,] 15232    NA   NA
[3,] 13375 15415   NA

$sequence
  X1    X2 X3
1  A     C NA
2  A <NA> NA
3  A     M NA

$path
[1] "HAC" "HA"  "HAM"

$ns
[1] 3 2 3
```

The Julian dates are converted back to calendar dates (class ‘Date’) using the `as.Date` function. The results is a data frame, which in the code is called `dates`.

The final step is the assemble the data in a data frame and to add the date format as an attribute. The following code produces the Biograph object (Table 3):

```
bio  <- data.frame (
          ID=id,
          born=born,
          start=start,
          end=interview,
          sex=sex,educ=educ,
          idim=as.numeric(rep(1,length(id))),
          ns=as.numeric(ns),
          path=as.character(path),
          dates[,1:(max(ns)-1)],stringsAsFactors=FALSE)

attr(bio,"format.date") <- "%Y-%m-%d"
```

Table 3. Biograph object. Hypothetical data A.

```
  ID       born      start        end sex   educ idim ns path        Tr1        Tr2
1  1 1986-04-05 1986-04-05 2019-05-09   F   High    1  3  HAC 2004-08-15 2011-12-15
2  2 1986-08-08 1986-08-08 2019-05-09   M Medium    1  2   HA 2011-09-15       <NA>
3  3 1986-11-28 1986-11-28 2019-05-09   F Medium    1  3  HAM 2006-08-15 2012-03-16
```

The data frame has different data types. The function `str(bio)` displays the data types (Table 4):

Table 4. Data types in Biograph object

```
'data.frame':	3 obs. of  11 variables:
 $ ID   : num  1 2 3
 $ born : Date, format: "1986-04-05" "1986-08-08" "1986-11-28"
 $ start: Date, format: "1986-04-05" "1986-08-08" "1986-11-28"
 $ end  : Date, format: "2019-05-09" "2019-05-09" "2019-05-09"
 $ sex  : Factor w/ 2 levels "F","M": 1 2 1
 $ educ : Factor w/ 2 levels "High","Medium": 1 2 2
 $ idim : num  1 1 1
 $ ns   : num  3 2 3
 $ path : chr  "HAC" "HA" "HAM"
 $ Tr1   : Date, format: "2004-08-15" "2011-09-15" "2006-08-15"
 $ Tr2   : Date, format: "2011-12-15" NA "2012-03-16"
 - attr(*, "format.date")= chr "%Y-%m-%d"
```

Note that the path variable must be a character variable. It should not be a factor variable. The covariates are factor variables.

The *Biograph* function Parameters can be invoked to check whether the Biograph object is correctly specified: `Parameters (bio).` The object produced by the function lists the states in the state space and identifies absorbing states. The latter are states that are entered but left during the observation period. It shows the lowest age and the highest age in the observation period. It also shows the transition matrix, which consists of logical values: a ‘TRUE’ indicates the transitions that occur during the observation period and a ‘FALSE’ identifies the transitions that do not occur during the observation period. It shows the line numbers of the transitions and the frequency of transitions `($nntrans)`. Finally, it lists the covariates and displays the date format. In this case the dates are of class ‘Date’ and a character string `"%Y-%m-%d"` gives the date format.

The R code for preparing the *Biograph* object that includes the data on the three subjects is given in the Annex.

Dates are often expressed in CMC. The preparation of a *Biograph* object requires the same procedure. Let’s convert the calendar dates to CMC, using the function `Date.as.cmc` of the *Biograph* package:

```
bio.cmc <- date.b (
              Bdata=bio,
              format.in="%Y-%m-%d",
              selectday=15,
              format.out="cmc",
              covs=NULL)
```

The Biograph object is shown in Table 5.

Table 5. Biograph object with dates in CMC

```
  ID born start  end sex   educ idim ns path  Tr1  Tr2
1  1 1036  1036 1433   F   High    1  3  HAC 1256 1344
2  2 1040  1040 1433   M Medium    1  2   HA 1341   NA
3  3 1043  1043 1433   F Medium    1  3  HAM 1280 1347
```

The R code to produce the *Biograph* object with dates in CMC is given in the Annex.

3. Hypothetical data B

Suppose we have information on a sample of 22 individuals. The state space is {H,A, B, C}, with C an absorbing state. Assume that the information is collected retrospectively as part of a cross-sectional survey. The date of interview is the end of the observation period. Since the data are collected retrospectively, no one drops out during observation. The respondents are born in 1991 and start in state H. The exact date of birth is unknown but it is assumed that births are uniformly distributed throughout the year. The date of birth is obtained by adding a random number between 0 and 365 to 1st January 1991. For each individual, seven dates are given: the date of birth, the date at entry into observation, the dates of the events under study with a maximum of three events (HA, AB and BC), and the date of interview. Of the 22 individuals, 10 do not experience a transition during the observation period, 4 experience one transition, 2 experience 2 transitions and 6 three. Respondent 1 is born on 31st July 1991 and enters observation on 2nd January 2007. He experiences the first event on 11th February, when he leaves the origin state (state 1) and enters the destination state (state 2). On 23rd March, he experiences the second event. That event implies a transition from state 2 to state 3. On 5th May he makes a transition to state 4. He stays in that state until the end of observation on 25th May.

The function `Sequences.ind.0` orders the dates chronologically and derives the state sequence. The components `f$d`, `f$path` and `f$ns` are included in the *Biograph object*. The following code produces the *Biograph* object:

```
RS <- data.frame (ID=id,
                  born=birth,
                  start=as.Date(entry,"%d/%m/%Y"),
                  end=as.Date(interview,"%d/%m/%Y"),
                  cov=cov,
                  idim=as.numeric(rep(1,length(id))),
                  ns=as.numeric(ns),
                  path=as.character(path),
                  dates[,1:(max(ns)-1)],
                  stringsAsFactors=FALSE)
attr(RS,"format.date") <- "%Y-%m-%d"
```

Table 6. Hypothetical survey data: multiple transitions

```
   ID       Born      Start       Stop          A          B          C
1   1 31/07/1991 02/01/2007 25/05/2007 11/02/2007 23/03/2007 05/05/2007
2   2 31/12/1991 17/01/2007 17/05/2007 04/05/2007         NA         NA
3   3 21/04/1991 18/01/2007 10/05/2007         NA         NA         NA
4   4 11/08/1991 22/01/2007 13/05/2007 28/02/2007 10/04/2007 10/05/2007
5   5 17/07/1991 10/02/2007 23/05/2007 17/05/2007         NA         NA
6   6 28/06/1991 30/01/2007 15/05/2007 12/02/2007 05/03/2007 17/04/2007
7   7 01/09/1991 04/04/2007 06/05/2007         NA         NA         NA
8   8 06/11/1991 29/04/2007 27/05/2007         NA         NA         NA
9   9 24/01/1991 18/05/2007 29/05/2007         NA         NA         NA
10 10 25/03/1991 20/05/2007 31/05/2007         NA         NA         NA
11 11 29/04/1991 15/05/2007 18/05/2007         NA         NA         NA
12 12 14/11/1991 05/02/2007 19/05/2007 25/02/2007 01/04/2007 02/05/2007
13 13 07/01/1991 05/02/2007 10/05/2007 18/04/2007 30/04/2007         NA
14 14 14/02/1991 06/02/2007 28/05/2007 18/05/2007 20/05/2007         NA
15 15 27/04/1991 26/02/2007 22/05/2007         NA         NA         NA
16 16 08/08/1991 10/03/2007 25/05/2007         NA         NA         NA
17 17 04/02/1991 11/03/2007 12/05/2007 08/05/2007         NA         NA
18 18 05/11/1991 28/03/2007 29/05/2007         NA         NA         NA
19 19 09/04/1991 15/03/2007 10/05/2007 23/03/2007 08/04/2007 20/04/2007
20 20 24/12/1991 13/04/2007 20/05/2007         NA         NA         NA
21 21 16/04/1991 04/04/2007 11/05/2007 09/05/2007         NA         NA
22 22 31/03/1991 25/04/2007 31/05/2007 16/05/2007 20/05/2007 26/05/2007
```

De Biograph object is shown in Table 7.

Table 7 Biograph object. Hypothetical data B

```
   ID       born      start        end cov idim ns path        Tr1        Tr2        Tr3
1   1 1991-12-18 2007-01-02 2007-05-25   X    1  4 HABC 2007-02-11 2007-03-23 2007-05-05
2   2 1991-12-28 2007-01-17 2007-05-17   X    1  2   HA 2007-05-04       <NA>       <NA>
3   3 1991-09-06 2007-01-18 2007-05-10   X    1  1    H       <NA>       <NA>       <NA>
4   4 1991-08-17 2007-01-22 2007-05-13   X    1  4 HABC 2007-02-28 2007-04-10 2007-05-10
5   5 1991-09-18 2007-02-10 2007-05-23   X    1  2   HA 2007-05-17       <NA>       <NA>
6   6 1991-11-20 2007-01-30 2007-05-15   X    1  4 HABC 2007-02-12 2007-03-05 2007-04-17
7   7 1991-05-07 2007-04-04 2007-05-06   X    1  1    H       <NA>       <NA>       <NA>
8   8 1991-12-09 2007-04-29 2007-05-27   X    1  1    H       <NA>       <NA>       <NA>
9   9 1991-02-28 2007-05-18 2007-05-29   X    1  1    H       <NA>       <NA>       <NA>
10 10 1991-05-26 2007-05-20 2007-05-31   X    1  1    H       <NA>       <NA>       <NA>
11 11 1991-07-11 2007-05-15 2007-05-18   X    1  1    H       <NA>       <NA>       <NA>
12 12 1991-05-29 2007-02-05 2007-05-19   X    1  4 HABC 2007-02-25 2007-04-01 2007-05-02
13 13 1991-09-18 2007-02-05 2007-05-10   X    1  3  HAB 2007-04-18 2007-04-30       <NA>
14 14 1991-12-06 2007-02-06 2007-05-28   X    1  3  HAB 2007-05-18 2007-05-20       <NA>
15 15 1991-11-14 2007-02-26 2007-05-22   X    1  1    H       <NA>       <NA>       <NA>
16 16 1991-01-22 2007-03-10 2007-05-25   X    1  1    H       <NA>       <NA>       <NA>
17 17 1991-12-11 2007-03-11 2007-05-12   X    1  2   HA 2007-05-08       <NA>       <NA>
18 18 1991-08-06 2007-03-28 2007-05-29   X    1  1    H       <NA>       <NA>       <NA>
19 19 1991-08-11 2007-03-15 2007-05-10   X    1  4 HABC 2007-03-23 2007-04-08 2007-04-20
20 20 1991-11-09 2007-04-13 2007-05-20   X    1  1    H       <NA>       <NA>       <NA>
21 21 1991-05-28 2007-04-04 2007-05-11   X    1  2   HA 2007-05-09       <NA>       <NA>
22 22 1991-11-11 2007-04-25 2007-05-31   X    1  4 HABC 2007-05-16 2007-05-20 2007-05-26
```

The data types in the data frame are shown in Table 8.

Table 8. Data types in Biograph object

```
'data.frame':     22 obs. of  11 variables:
 $ ID   : int  1 2 3 4 5 6 7 8 9 10 ...
 $ born : Date, format: "1991-12-18" "1991-12-28" "1991-09-06" ...
 $ start: Date, format: "2007-01-02" "2007-01-17" "2007-01-18" ...
 $ end  : Date, format: "2007-05-25" "2007-05-17" "2007-05-10" ...
 $ cov  : chr  "X" "X" "X" "X" ...
 $ idim : num  1 1 1 1 1 1 1 1 1 1 ...
 $ ns   : num  4 2 1 4 2 4 1 1 1 1 ...
 $ path : chr  "HABC" "HA" "H" "HABC" ...
 $ Tr1  : Date, format: "2007-02-11" "2007-05-04" NA ...
 $ Tr2  : Date, format: "2007-03-23" NA NA ...
 $ Tr3  : Date, format: "2007-05-05" NA NA ...
 - attr(*, "format.date")= chr "%Y-%m-%d"
 - attr(*, "trans")= num [1:4, 1:4] NA NA NA NA 1 NA NA NA NA 2 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ From: chr [1:4(1d)] "H" "A" "B" "C"
  .. ..$ To  : chr [1:4(1d)] "H" "A" "B" "C"
```

4. GLHS data

In this section, I describe how to prepare a Biograph object from the German Life History Survey (GLHS) data published by Blossfeld and Rohwer (2002).

The German Life History Survey (GLHS) provides detailed retrospective information on life histories of more than 5,000 men and women from three birth cohorts: 1929-31, 1939-41, and 1949-51. The data are collected in the years 1981-83. A subsample of the GLHS is used for training purposes by Blossfeld and Rohwer (2002) and Blossfeld et al. (2007). It consists of 201 respondents with 600 job episodes. Blossfeld and co-authors illustrate two software packages: TDA (Transition Data Analysis) in the 2002 publication and Stata in the 2007 publication. The data file (rrdat.1) can be downloaded from www.soziologie-blossfeld.de/eha/tda/index.html. The authors study the 600 job episodes. *Biograph* considers the full employment career that includes 600 job spells and 382 spells without a job. It addresses the complete sequence of both states and events that characterise the employment domain of the life course.

A selection of the GLHS survey data is presented in Table 9. The rrdat.1 file is an episode file with one record for each job episode (long format). The data shown in Table 9 is the format that the GLHS data provided by Blossfeld and Rohwer come in. The data contain the date of birth and 5 covariates: sex, date of marriage, prestige score of the current job, prestige score of the next job and level of education. Education is the years of education derived from the highest educational attainment before entry into the labour market (Blossfeld and Rohwer, 2002, p. 44). Lower secondary school qualification without vocational training is equivalent to 9 years, middle school qualification 10 years, lower secondary school with vocational training 11 years, middle school with vocational training 12 years, Arbitur 13 years, professional college qualification 17 years and university degree 19 years. Observation starts at birth (TB) and ends at the date of interview (TI). A job episode is characterised by the serial number of the job episode (NOJ), the starting date of the episode (TS) and the ending date (TF). The starting date of the first job episode is the date of entry into the labour market. Dates are given in Century Month Code (CMC).

Consider subject 1. He is a male (sex = 1) born at CMC 351 (March 1929). He enters the first job at CMC 555 (March 1946). That first job episode ends at survey date in CMC 982 (October 1981). The birth cohort is 1 (1929-31). The respondent enters the second episode, which is a job episode, at CMC 555. He leaves observation at CMC 982.

Table 9 TDA input data file rrdat: episode file

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
1	1	555	982	1	982	351	555	679	34	-1	17
2	1	593	638	2	982	357	593	762	22	46	10
2	2	639	672	2	982	357	593	762	46	46	10
2	3	673	892	2		357	593	762	46	-1	10

Variable	Name	Description
1	ID	Identification number of subject
2	NOJ	Serial number of the job episode
3	TS	Starting time of the job episode
4	TF	Ending time of the job episode
5	SEX	Sex (1 male; 2 female)
6	TI	Date of interview (CMC)
7	TB	Date of birth (CMC)
8	T1	Date of entry into the labour market (CMC) (denoted by TE)
9	TM	Date of marriage (CMC) [0 if not married]
10	PRES	Prestige score of current job, i.e. of job episode in current record of data file
11	PRESN	Prestige score of the next job (if missing: -1)
12	EDU	Highest educational attainment before entry into labour market

To convert the Blossfeld-Rohwer (2002) data into a *Biograph* object, two steps are required. First is to read the data from the designated website http://oldsite.soziologie-blossfeld.de/eha/tda/ and to create a data frame. The name of the data frame is rrdat. The code is

```
url.tda <- "http://oldsite.soziologie-
 Blossfeld.de/eha/tda/cf_files/Data/RRDAT.1"
rrdat.1 <- as.matrix (read.table(file=url.tda),header=FALSE)
colnames(rrdat.1) <- c("ID","NOJ","TS","TF","SEX","TI","TB","TE",
   "TM","PRES","PRES1","EDU")
rownames(rrdat.1) <-c(1:nrow(rrdat.1))
rrdat <- data.frame(rrdat.1)
```

In the second step, the *Biograph* object is prepared, following the five steps described in the previous sections. The state space consists of two states: no job (N) and job (J). Everyone starts in state N. The `reshape` function is used to convert the long format to a wide format. When creating the wide format, the attributes of episodes (NOJ, PRES and PRESN) are omitted and a new covariate (birth cohort) is defined.
The Blossfeld-Rohwer data are limited to job episodes, with information on the starting month and ending month of a job episode. Note that Blossfeld and Rohwer assume that job episodes start at the beginning of the month and end at the end of the month. In *Biograph*, the end of an episode is not considered explicitly because the end of an episode is the beginning of a new episode. Episodes are assumed to start at the

beginning of the month. From that data on job episodes, the start and end of episodes without a job are extracted.

Table 10. Biograph object. Selection of GLHS data.

```
  ID born start end    sex edu marriage LMentry  cohort idim ns     path Tr1 Tr2 Tr3 Tr4 Tr5 Tr6 Tr7
1  1  351   351 983   Male  17      679     555 1929-31    1  2       NJ 555  NA  NA  NA  NA  NA  NA
2  2  357   357 983 Female  10      762     593 1929-31    1  5    NJJJN 593 639 673 893  NA  NA  NA
3  3  473   473 983 Female  11      870     688 1939-41    1  7  NJJJJJN 688 700 730 742 817 829  NA
4  4  604   604 983 Female  13      872     872 1949-51    1  3      NJN 872 927  NA  NA  NA  NA  NA
5  5  377   377 983   Male  11      701     583 1929-31    1  4     NJJJ 583 651 788  NA  NA  NA  NA
6  6  492   492 983   Male  11      781     691 1939-41    1  8 NJNJNJNJ 691 717 728 754 771 847 859
```

Two attributes are added to the data set. The first is the format of the transition dates:

```
attr(GLHS,"format.date") <- "CMC"
```

The second is

```
z <- Parameters (GLHS)
attr(GLHS,"trans") <- z$tmat
```

The *Biograph* object is now complete. By way of illustration, the TRANS function is invoked: `Trans(GLHS)`. The number of transitions and censoring are shown in Table 11.

Table 11. Number of transitions and censoring. GLHS.

```
          Destination
Origin      N   J Total  Censored
  N         0 323   323        59
  J       181 277   458       142
  Total   181 600   781       201
```

5. Netherlands Family and Fertility Survey 1998 (NLOG98)

Between February and May 1998 Statistics Netherlands (CBS) conducted the Netherlands Family and Fertility Survey (NLOG98). Data were collected on 5,450 women and 4,717 men, born in the period 1945-79 and residing in The Netherlands. They were 18 to 52 years at time of survey. The sample frame consisted of the Municipal Population Administration (Gemeentelijke Bevolkingsadministratie; GBA). The GBA is the main source of statistical information on the population of the Netherlands. The random sample survey was organized in two steps. In the first step 262 municipalities were selected from the 572 municipalities. The GBA of the selected municipalities was used to randomly select 14 thousand addresses and subsequently men and women born in the period of 1945-79. The drawing of the random sample was done taking into account several conditions (for details refer to de Graaf and Steenhof, 1999, p. 36). Eventually, 5450 women and 4717 men were interviewed using structured questionnaires. In this chapter, we use data on women only.

The NLOG98 provides extensive information on marital status, living arrangements, partnership and fertility. For each respondent, the OG98 reports up to three marriages and up to six cohabitations. Each marriage may be followed by a divorce or widowhood.

The data may be obtained from DANS (Data Archiving and Networked Services) in The Hague (http://www.dans.knaw.nl/). The data are distributed in two SPSS files. The file BOAV98.SAV contains the data for females and the file BOAM98.SAV contains the data for males. I use the data on females.

The raw data need some processing to be useful for the study of life histories. First, the public use file of NLOG98 does not include the survey month. Although we know that the survey took place from February to May 1998, the month of interview is not given and is not available to researchers. The age of the respondent at the time of survey is given, however. The CMC at survey is estimated from the age at survey and the month of birth of the respondent. The estimation procedure includes a random number generation to allocate the survey date to one of several plausible months. Second, the public use data file is not well suited for life history data analysis. The focus of the questionnaire was on partnership and not on timing of events. Life history data analysis requires that the events are ordered and defined in terms of origin state, destination state and date of occurrence. The conversion of raw data into an event history data structure is a tedious process that was completed by Matsuo and Willekens (2003). The dates of events are recoded in century month codes (CMC). If necessary, imputation is accomplished. The emphasis on the sequence and timing of events did reveal several inconsistencies in the data that remain hidden otherwise. Particular sequences of events may not be possible (e.g. second child is born before first child) or plausible (e.g. marriage before leaving the parental home). Events may be missing (e.g. second marriage is reported while information on dissolution of first marriage is missing). The inconsistencies were investigated in detail and corrected if it was clear that the inconsistent sequence or timing of events was due to errors in recording or coding. The report by Matsuo and Willekens (2003) and a set of about 10 SPSS syntax files that convert the original Public Use Data File into an event history data structure are available from the website of the Population Research Centre (PRC), University of Groningen (accessed 25 March 2012):

http://www.rug.nl/prc/publications/researchReports/index

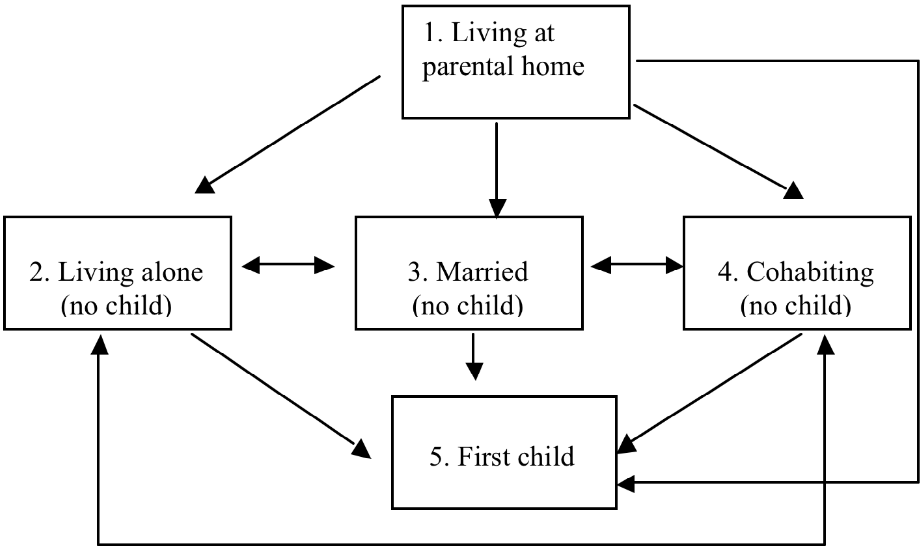
The last SPSS syntax file saves the relevant data in the event history file. The name of the event history file is NLOG98_F_CMC.sav.

The life path starts with the state of living at the parental home. We assume that the parental home may be left only once, although in reality persons may leave the parental home and return later at least for some time. The respondent may leave home for one of three reasons. The first is independence, which is manifested by leaving home to live alone. The second and third reasons involve union formation through marriage (second reason) or cohabitation (third reason). Childbearing may occur in any of the states. The states are:

- Living at parental home (H)
- Living alone (independently) (A)
- Married (M)
- Cohabiting (C)
- First child (K)

The state space is determined by a composite variable that combines three domains of life. The first domain of life is the living arrangement with three possibilities: living at the parental home, living alone, and living with someone. The second domain of life is the marital status: not married or married. The third domain is motherhood (fertility). The three state spaces are combined into a single state space and some combinations of states are excluded (e.g. cohabitating at the parental home, married while living at the parental home). Figure 1 shows the transitions considered in the analysis.

Figure 1. Schematic representation of pathways to first child



The following variables are extracted from the data file NLOG98_F_CMC.sav:

Dates:

Variable name	*Meaning*
- CMCINT	CMC at interview
- CMCB_OP	CMC at birth
- CMCLEAVE	CMC at leaving parental home
- CMCCO1	CMC at first cohabitation
- CMCE1CO	CMC at end first cohabitation
- CMCCO2	CMC at second cohabitation
- CMCE2CO	CMC at end second cohabitation
- CMCCO3	CMC at third cohabitation
- CMCE3CO	CMC at end third cohabitation
- CMCCO4	CMC at fourth cohabitation
- CMCCO5	CMC at fifth cohabitation
- CMCMA1	CMC at first marriage
- CMCE1MA	CMC at end of first marriage
- CMCMA2	CMC at second marriage
- CMCE2MA	CMC at end of second marriage
- CMCMA3	CMC at third marriage

- CMCE3MA	CMC at end of third marriage
- CMC_K1	CMC at birth of first child

Covariates:

Variable name	*Meaning*
- kerkgez	Religion
- educ	Highest completed education

Two covariates are included in the data. The first is religion (labeled kerkgez in the original data distributed by Statistics Netherlands). The following categories are distinguished:

1. No religion (code 1 in original data file): 2395
2. Roman Catholic (code 2 in original data file): 1677
3. Protestant (codes 3, 4, 5 and 6 in original data file): 1014
4. Other religion (codes 7, 8, 9 and 10 in original data file): 357

NA Missing data (coded 98, 99 and missing in original data file): 7

The second is the highest completed education. The following categories are distinguished:

1. Primary (code 2 in original data file): 363
2. Secondary lower (code 3 in original data file): 1250
3. Secondary higher (code 4 in original data file): 2489
4. First step high (code 5 in original data file): 869
5. Second step high (code 6 in original data file): 238
6. Third step high (code 20 in original data file): 20

NA Missing data (code 9 and missing in original data file): 221

In addition, two birth cohorts are derived from the dates of birth. A total of 2306 respondents are born before 1960 and 3144 are born in 1960 or later.

A number of respondents started a new cohabitation or marriage in the month the previous cohabitation or marriage ended. I assume that the transition was direct and not via a period of living alone. If a respondent started cohabitation and marriage in the month, I assume that the transition is to marriage.

The interview date is given in CMC and the assumption is made that interview is at the end of the month (estimated). Since *Biograph* assumes that events, including censoring, is at the beginning of a month, a one is added to the interview month.

A data frame of transition dates is constructed and the NLOG98 variable labels (transitions) are replaced by the labels of **destination** states used in *Biograph*. The function `Sequences.ind.0` orders the dates chronologically and derives the state sequence. The components `f$d`, `f$path` and `f$ns` are included in the *Biograph* object. Table 12 shows a selection of the data in *Biograph* format.

Two attributes are added to the object. The first is the format of the transitions dates (cmc). The second is the transition matrix, i.e. the matrix of possible transitions. The matrix is:

```
                    To
          From  H  C  A  M  K
             H NA  1 NA  2  3
             C NA  4  5  6  7
             A NA  8  9 10 11
             M NA 12 13 NA 14
             K NA NA NA NA NA
```

Table 12. Biograph object. NLOG98

```
   ID born start  end          kerk educ cohort YearInt idim ns  path  Tr1  Tr2  Tr3  Tr4
4   4  787   787 1180    no religion    3  1960+    1998    1  2    HC 1159   NA   NA   NA
5   5  577   577 1179    no religion   NA  <1960    1998    1  1     H   NA   NA   NA   NA
6   6  734   734 1182    no religion    4  1960+    1998    1  4  HCAC  979  981 1078   NA
7   7  591   591 1181    no religion    6  <1960    1998    1  5 HCACM  882  986 1003 1010
8   8  707   707 1180 Roman Catholic   NA  <1960    1998    1  4  HCMK  894  906  910   NA
9   9  661   661 1179    no religion    4  <1960    1998    1  5 HCACK  889  973 1059 1059
10 10  571   571 1179    no religion    4  <1960    1998    1  3   HMK  816  828   NA   NA
```

The function `Trans (NLOG98)` produces numbers of transitions and censoring (Table 13).

Table 13. Number of transitions and censoring. NLOG98

```
Destination
Origin  H    C   A    M    K Total Censored
  H     0 2363   0 2105  110  4578      872
  C     0   44 506 1460  258  2268      566
  A     0  392   1   46   32   471      209
  M     0   35 173    0 2996  3204      407
  K     0    0   0    0    0     0     3396
  Total 0 2834 680 3611 3396 10521     5450
```

6. Survey of Health, Ageing and Retirement in Europe (SHARE)

The Survey of Health, Ageing and Retirement in Europe (SHARE) (http://www.share-project.org/) is a multidisciplinary and cross-national panel database of micro data on health, socio-economic status and social and family networks of more than 55,000 individuals aged 50 or over from 20 European countries. SHARE is harmonized with the U.S. Health and Retirement Study (HRS) and the English Longitudinal Study of Ageing (ELSA). The SHARE baseline study (wave 1) was carried out in 2004. The third wave of data collection for SHARE (2008-09) focused on people's life histories. It is referred to as SHARELIFE. Almost 30,000 men and women across 13 European countries took part in this round of the survey. The respondents are representative for the European population aged 50 and over in Scandinavia (Denmark and Sweden), Central Europe (Austria, France, Germany, Switzerland, Belgium, and the Netherlands), and the Mediterranean (Spain, Italy and Greece), as well as two transition countries (the Czech Republic and Poland). The SHARELIFE questionnaire covers different domains of life, ranging from partners and children over housing and work history to detailed questions on health and health care. The SHARELIFE questionnaire has several modules and the data from each module are stored in a different data file. The following modules and data files are distinguished:

ac	Accommodation section
cs	Childhood section
dq	Disability
fs	financial history
gl	General life questions
gs	Grip strength
hc	Childhood health care
hs	Childhood health section
iv	Interviewer
rc	Retrospective children
re	Work history
rp	Partner section
st	Demographics
wq	Work quality
xt	End of life interview

The data are available for download after registration. Applicants must have a scientific affiliation and have to sign a statement confirming that under no circumstances the data will be used for other than purely scientific purposes. Data are available as SPSS and STATA files.

For the illustration of *Biograph*, I selected data on partnerships and living arrangement and downloaded the STATA files. The code to read the downloaded data is:

```
d.st <- data.frame(read.dta
 ("sharew3_rel1_st.dta",convert.dates=TRUE,convert.underscore=TRUE))
d.rp <- data.frame(read.dta
 ("sharew3_rel1_rp.dta",convert.dates=TRUE,convert.underscore=TRUE))
d.ac <- data.frame(read.dta
 ("sharew3_rel1_ac.dta",convert.dates=TRUE,convert.underscore=TRUE))
d.re <- data.frame(read.dta ("  /
 sharew3_rel1_re.dta",convert.dates=TRUE,convert.underscore=TRUE))
d.rc <- data.frame(read.dta
("sharew3_rel1_rc.dta",convert.dates=TRUE,convert.underscore=TRUE))
```

The state space is:

- Living at parental home (H)
- Living alone (independently) (A)
- Cohabiting (C)
- Married (M)

Four covariates are considered:

- Sex
- Education: year in which full-time education is finished
- Number of children
- Year in which the respondent started his/her first job

In addition four birth cohorts are determined: `<1930, 1930-39, 1940-49, 1950+`

The following variables are extracted from the data:

Dates:

Variable name	*Meaning*
d.st$mergeid	Identification number
d.st$sl.st007	Year of birth
d.st$sl.st006	Month of birth
d.ac$sl.ac003.	Year of leaving parental home
d.rp$sl.rp008.1	Year of first marriage
d.rp$sl.rp008.k	Year of k-th marriage (k = 1 to 6)
d.rp$sl.rp013.k	Divorce (k – 1 to 4) (yes/no)
d.rp$sl.rp014.k	Year of k-th divorce (k = 1 to 4)
d.rp$sl.rp004b.k	Year in which k-th cohabitation before a marriage started (k = 1 to 6)
d.rp$sl.rp012.k	Year in which k-th cohabitation ended (k = 1 to 4)
d.rp$sl.rp003.n	Year in cohabitation NOT related to marriage started (n = 11 to 18)
d.rp$sl.rp012.n	Year in which cohabitation ended

Covariates:

Variable name	*Meaning*
d.st$sl.st011.	Sex
d.re$sl.re002.	Year in which full-time education is finished
d.rc$sl.rc023.	Number of children
d.re$sl.re011.1	Year of entry in labour market

A data frame of transition dates is constructed and the SHARELIFE variable labels (transitions) are replaced by the labels of destination states used in *Biograph*. The next step is to sort the dates at transition, using the `Sequence.ind.0` function. The function produces state sequences and the sequence of dates at transition. The following code stores the data in a data frame:

```
SHARE<- data.frame(ID=c(1:nsample),
                   born=as.numeric(bb),
                   start=as.numeric(bb),
                   end=as.numeric(end),
                   country=as.factor(d.st$country),
                   IDc=IDc,
                   cohort=bcohort,
                   sex=as.factor(sex2),
                   eduf=as.numeric(edu.f),j
                   ob1=as.numeric(job.1.start),
                   children=nchildren,
                   idim=as.numeric(rep(1,length(IDc))),
                   ns=as.numeric(ns),
                   path=as.character(path),
                   f$d[,1:(max(ns)-1)])
```

Two attributes are added to the data file: the format of the transition dates (year) and the transition matrix. For one respondent the date of birth is missing. The row with the data on that respondent is removed.

Table 14 shows a selection of rows of the SHARELIFE data in the *Biograph* format. Table 15 shows numbers of transitions and censoring.

Table 15. Number of transitions and censoring. SHARELIFE

```
              Destination
Origin  H     M     A    C Total Censored
  H     0 17941  7847  595 26383      343
  M     0   303  5066  882  6251    21126
  A     0  5023  1638 6251 12912     3716
  C     0  4110  2077  220  6407     1441
  Total 0 27377 16628 7948 51953    26626
```

7. National Family Health Survey of India 2005-06 (NFHS): Andhra Pradesh

The National Family Health Survey (NFHS) (http://www.nfhsindia.org/) is a large-scale, multi-round survey conducted in a representative sample of households throughout India. In total 109,041 households were interviewed. The survey provides state and national information for India on fertility, infant and child mortality, the practice of family planning, maternal and child health, reproductive health, nutrition, anaemia, utilization and quality of health and family planning services. NFHS surveys are conducted under the stewardship of the Ministry of Health and Family Welfare (MOHFW), Government of India. The nodal agency, responsible for coordination and technical guidance is the International Institute for Population Sciences (IIPS) in Mumbai.

Three rounds of the survey have been conducted since the first survey in 1992-93. The second survey was organized in 1998-99 and the third in 2005-06. The third survey (NFHS-3) covered all 29 states in India, which comprise more than 99 percent of India's population. The survey included 124,385 women and 74,369 men with completed interview (married and unmarried). Women interviewed were between ages 15 and 49, while men were between 15 and 54. All dates are in Century Month Code (CMC).

The data are available for download (after registration) through the Demographic and Health Survey (DHS) data distribution system (http://www.measuredhs.com). Data files are available in user-friendly formats for SPSS, SAS, and STATA users. For the illustration of *Biograph*, I used the SPSS data file named APIR42RT.SAV and more particularly the data for women from the state of Andhra Pradesh (AP). The survey covered 5,153 women. The number of variables is 4,386. For the main survey report, see IIPS and Macro International (2007).

Suppose we are interested in the fertility career of women: when they marry, whether and when they have children, and whether and when they opt for sterilization. The state space is:

- Never married (N)
- Married without children (M)
- One child (a)
- Two children (b)
- Three children (c) up to 20 children (m)
- Sterilized (S)

The following variables are extracted from the raw data:

Dates:

Variable name	*Meaning*
v011	Date of birth
v008	Date of interview
v509	Date of first marriage
b3.*	Date of birth of child (from youngest to oldest)
bord.*	Birth order of child
v312	Contraceptive method (sterilization = 6 (female) or 7 (male))
v317	Date of sterilization

Covariates:

Variable name	*Meaning*
v106	Level of education
v190	Wealth index
v102	Place of residence (urban/rural)
v201	Number of children ever born (nCEB)

In addition, three birth cohorts (COH) are distinguished: born before 1970, between 1970 and 1979, and in 1980 or later.

The observation window starts at birth and ends at time of interview. The date of interview is given in CMC. I assume that interview takes place at the beginning of the month.

The raw data presents the months of birth of the children starting with the youngest child. In *Biograph* the dates should be ordered chronologically, i.e. from the birth of the oldest child to the birth of the youngest and last child. The first step is to arrange the CMCs at birth of children from the oldest child to the youngest child. The result is the object `cmc_k06`. The CMC at first marriage and the CMC at sterilization of the woman or her spouse are added next. A missing value (NA) indicates the absence of sterilization. The dates are stored in the data frame `cmc`. The next step is to sort the dates at transitions, using the standard `Sequence.ind.0` function. The function produces state sequences and the sequence of dates at transition.

The data are stored in a data frame (AP). Table 16 shows a selection of rows.

Table 16 Biograph object. NFHS-AP

```
   ID born start  end      COH EDU WEAL U_R idim ns     path  Ev1  Ev2  Ev3  Ev4  Ev5  Ev6
1   1  709   709 1273   <1970   0    2   2    1  7  HMabcdS  936  937  964 1006 1045 1045
2   2  997   997 1273  >=1980   1    2   2    1  5    HMabS 1200 1210 1238 1238   NA   NA
3   3 1033  1033 1275  >=1980   0    2   2    1  3      HMa 1172 1197   NA   NA   NA   NA
4   4 1009  1009 1273  >=1980   0    3   2    1  5    HMabS 1193 1202 1221 1221   NA   NA
5   5  973   973 1273  >=1980   2    3   2    1  5    HMabS 1169 1200 1211 1211   NA   NA
6   6  733   733 1273   <1970   0    4   2    1  6   HMabcS  919  949  997 1040 1046   NA
7   7  985   985 1273  >=1980   2    4   2    1  5    HMabS 1241 1250 1262 1263   NA   NA
8   8 1011  1011 1273  >=1980   0    3   2    1  3      HMa 1205 1238   NA   NA   NA   NA
```

The function `Trans(AP)` is used to produce the table of numbers of transitions and censored cases by state at time of censoring (Table 17). Of the 5,153 respondents, 4,603 have at least one child at time or survey and 9 have at 10 children or more. Among the respondents are 3,209 women who are sterilized or have a husband who is sterilized. One childless woman and 55 women with one child went for sterilization. Most couples in Andhra Pradesh who opt for sterilization have the operation after two or three children.

Table 17. Number of transitions and censoring. NFHS-AP

```
                                 Destination
Origin  H    M    a    b    c    d    S   e   f  g  h  i j k l m Total Censored
  H     0 5099   54    0    0    0    0   0   0  0  0  0 0 0 0 0  5153        0
  M     0    0 4548   33    9    4    1   2   0  0  0  0 0 0 0 0  4597      556
  a     0   35    0 3852    0    0   55   0   0  0  0  0 0 0 0 0  3942      661
  b     0   10    0    0 2184    0 1276   0   0  0  0  0 0 0 0 0  3470      416
  c     0    4    0    0    0 1041 1027   0   0  0  0  0 0 0 0 0  2072      123
  d     0    2    0    0    0    0  525 433   0  0  0  0 0 0 0 0   960       88
  S     0    2    1    1    2    3    0   1   0  0  0  0 0 0 0 0    10     3199
  e     0    0    0    0    0    0  183   0 207  0  0  0 0 0 0 0   390       46
  f     0    0    0    0    0    0   84   0   0 97  0  0 0 0 0 0   181       26
  g     0    0    0    0    0    0   39   0   0  0 46  0 0 0 0 0    85       12
  h     0    0    0    0    0    0   11   0   0  0  0 21 0 0 0 0    32       14
  i     0    1    0    0    0    0    5   0   0  0  0  0 9 0 0 0    15        6
  j     0    0    0    0    0    0    1   0   0  0  0  0 0 4 0 0     5        4
  k     0    0    0    0    0    0    1   0   0  0  0  0 0 0 1 0     2        2
  l     0    0    0    0    0    0    0   0   0  0  0  0 0 0 0 1     1        0
  m     0    0    0    0    0    0    1   0   0  0  0  0 0 0 0 0     1        0
  Total 0 5153 4603 3886 2195 1048 3209 436 207 97 46 21 9 4 1 1 20916     5153
```

8. European Registry for Blood and Marrow Transplantation (EBMT)

a. Introduction

Bone marrow in large bones produces new blood cells (by the hematopoietic compartment of bone marrow where hematopoietic stem cells produce new blood cells). Bone marrow transplantation is a treatment for people with certain forms of cancer such as leukemia and lymphoma. High doses of chemotherapy or radiation therapy can effectively kill cancer cells but they also destroy bone marrow, where blood cells are made. The purpose of a bone marrow transplant is to replenish the body with healthy bone marrow after a high-dose chemotherapy or radiation therapy. Transplanted cells are able to rebuild the patient's bone marrow. After a successful transplant, the bone marrow will start to produce new blood cells. Bone marrow transplantation is also a treatment of acute leukemia patients whose bone marrow contains malignant cells.

The goal of cancer therapy is to bring the disease into remission. Remission is when the patient's blood counts return to normal and (in case of leukemia) bone marrow samples show no sign of disease. Patients may fail to attain a complete remission (CR) because of drug resistance or death. A percentage of patients who initially attain a CR will relapse. Relapse is the reoccurrence of the cancer. If the doses of therapy are not sufficiently high, they are not generally curative. They induce remission but the patient usually relapses. The purpose of bone marrow transplants is to provide the patient with healthy marrow so as to allow massive, and hopefully, curative doses of therapy.

There are two types of bone marrow transplants:

- *Autologous bone marrow transplant* - The donor of the bone marrow (hematopoietic stem cells) is the person him/herself.
- *Allogenic bone marrow transplant* - The donor is another person whose tissue has the same genetic type as the person needing the transplant (recipient). Because tissue types are inherited, it is more likely that the patient's brother or sister are suitable donors. If a family member does not match the recipient, the Marrow Donor Program Registry database is searched for an unrelated individual whose tissue type is a close match. If donor and recipient are compatible, the infused cells will then travel to the bone marrow and initiate blood cell production.

The European Group for Blood and Marrow Transplantation (EBMT) (http://www.ebmt.org/) maintains a patient database known as the EBMT Registry. The Registry goes back to the beginning of the 1970's and contains patient clinical data. The population covered are patients who have undergone an haematopoietic stem cell transplantation (HSCT) procedure; patients with bone marrow failures receiving immunosuppressive therapies; and patients receiving non-haematopoietic cell therapies. Patients are followed up indefinitely. The data base has data on close to 400 thousand patients. The data cover aspects of the diagnosis, first line treatments, HSCT (hematopoietic stem cell transplantation) or cell therapy associated procedures, complications and outcome. The transplant data are submitted to the central registry by EBMT member centres performing any of the above treatments. The purpose of the Registry is to provide a pool of data to perform retrospective studies, assess epidemiological trends, or prepare prospective trials.

b. The data

The *mstate* package, developed by Putter and colleagues (see de Wreede et al, 2011), includes part of the EBMT data. The data, in a file names `ebmt4`, are from 2279 acute lymphoid leukemia (ALL) patients who had an allogeneic bone marrow transplant from an HLA-identical sibling donor between 1985 and 1998. An HLA-identical donor is a donor who shares the same **H**uman**L**eukocyte **A**ntigens (HLA). The data were extracted from the EBMT database in 2004. All patients were transplanted in first complete remission. Events recorded during the follow-up of these patients were: acute graft versus host disease (AGvHD), platelet recovery (PR, the recovery of platelet counts to a level of >20°x109/l), relapse and death. AGvHD has been defined as a GvHD of grade 2 or higher, appearing before 100 days post-transplant.

Four prognostic factors are known at baseline for all patients. They are: donor-recipient gender match (where gender mismatch is defined as female donor, male recipient), prophylaxis, year of transplant and age at transplant in years. All these covariates are treated as time-fixed categorical covariates. Younger patients have a better prognosis and transplantation before 1990 had a worse prognosis. Donor recipient gender mismatch seems to be of minor importance, while TCD shows a clear negative effect on failure-free survival.

The data were used in Fiocco, Putter & van Houwelingen (2008) and van Houwelingen & Putter (2008). The included variables are

id	Patient identification number
Rec	Time in days from transplantation to recovery or last follow-up
rec.s	Recovery status; 1 = recovery, 0 = censored
ae	Time in days from transplantation to adverse event (AE) or last follow-up
ae.s	Adverse event status; 1 = adverse event, 0 = censored
recae	Time in days from transplantation to both recovery and AE or last follow-up
plag.s	Recovery and AE status; 1 = both recovery and AE, 0 = no recovery or no AE or censored
rel	Time in days from transplantation to relapse or last follow-up
rel.s	Relapse status; 1 = relapse, 0 = censored
srv	Time in days from transplantation to death or last follow-up
srv.s	Relapse status; 1 = dead, 0 = censored
year	Year of transplantation; factor with levels "1985-1989", "1990-1994", "1995-1998"
agecl	Patient age at transplant; factor with levels "<=20", "20-40", ">40"
proph	Prophylaxis; factor with levels "no", "yes"
match	Donor-recipient gender match; factor with levels "no gender mismatch", "gender mismatch"

c. The model

In their research, the authors opt for a multistate approach because it enables the distinction between disease-related and the treatment-related morbidity and mortality. Information on the occurrence of the intermediate events recovery and adverse event is used to update the prognoses of the patients. The multistate model considers six states (with the state labels used in *mstate* and *Biograph* in parentheses):

- Alive and in remission, no recovery or adverse event (Tx, T);
- Alive in remission, recovered from the treatment (Rec, P);
- Alive in remission, occurrence of the adverse event (AE, A);
- Alive, both recovered and adverse event occurred (Rec+AE, Z);
- Alive, in relapse (treatment failure) (Rel, R);
- Dead (treatment failure) (Death, D).

All patients start in state Tx. States Rel and Death are called absorbing: once the patient has entered one of them, she/he stays there. This leaves us with a model with 12 transitions. Time is measured in days since transplant. Status variables (.s) indicate the (non)occurrence of a transition. For instance patient 2 experienced the adverse event after 12 days (transition from state Tx to state AE), then recovery after 29 days (transition from state AE to state "Rec+AE") and a relapse after 422 days (transition from state "Rec+AE" to state Rel). Finally, he/she died after 579 days. The last event is not relevant to the model because the patient had already reached an absorbing state.

A few adjustments of the data were required for a multi-state analysis. Since the model does not allow patients to enter two states at the same time, a patient who experienced relapse and death at the same time is assumed to have entered the absorbing state of relapse rather than death, because the patients must have experienced the relapse before their death. Patients who experienced the adverse event and recovery at the same time are assumed to have experienced the AE half a day before Rec. Two new variables have been created to express the time of entry in state "Rec+AE" and the accompanying status indicator: recae and recae.s respectively.

For modeling, the events relapse and death are combined into a single event 'failure'. Three intermediate events are included in the model: Recovery (Rec), an Adverse Event (AE) and a combination of the two (AE and Rec). To avoid misinterpretation, the authors have abstracted from the actual disease, covariate values and intermediate events. Instead of Recovery, Engraftment or platelet recovery can be included. A platelet is a particle in the blood that is an important part of blood clotting. The bone marrow produces a large number of platelets per mm^3 of blood daily. During chemotherapy, the platelet count drops significantly. Engraftment is the process of transplanted stem cells reproducing new cells. The Adverse Event may be replaced by Acute Graft-versus-Host Disease (AGVHD). It is a complication that can occur after a bone marrow transplant in which the newly transplanted material attacks the transplant recipient's body. The data include four covariates: year at transplantation, age at transplantation,

d. Preparation of *Biograph* object

The preparation of a *Biograph* object involves the five steps listed in previous sections of the paper. The state space includes the six states shown above. It is {T, P, A, Z, R, D}. All patients start in state T. In *Biograph*, transitions are specified a little different from the specification of transitions in the data (`ebmt4`). In case an event occurs, both a *mstate* object and a *Biograph* object show the date of the event. In case an event does not occur, the *mstate* object lists the date at censoring, which is the end of exposure to the risk of experiencing that event. A *Biograph* object shows NA for not applicable. The preparation of a *Biograph* object involves the removal of censoring dates in cases of non-occurrence of transitions. Note that in *Biograph*, a transition is defined by the state of destination. The transition dates are stored in the data frame `days`. Table 18 shows the first rows of the data frame. The maximum number of transitions patients experience is 3.

The first patient recovered 22 days after transplantation. The second patient experienced an adverse event 12 days after transplantation, recovered at 29 days and experienced a relapse 422 days after transplantation. Patient 4 enters relapse 84 days after transplantation. The observation ends at that time.

Table 18. Data frame with event dates in days since transplantation

```
    P    A  Z   R   D
1  22   NA NA  NA  NA
2  NA 12.0 29 422  NA
3  NA 27.0 NA  NA  NA
4  NA 42.0 50  84  NA
5  22   NA NA 114  NA
```

The covariates are

	Variable name	*Meaning*
•	match	Donor-recipient gender match
•	proph	Prophylaxis
•	year	Year of transplantation
•	agecl	Patient age at transplant

The observation window is from date of transplantation to date of entry into the absorbing state. Time is measured in days since transplantation. The event dates are arranged chronologically and the state sequence determined by the function `Sequences.ind.0`:

```
f<- Sequences.ind.0 (days,namstates,absorb=c("R","D"))
```

Note the two absorbing states. The output component `f$path` gives for each patient the state sequences. The event dates in days since transplantation are given in `f$d`.

The data frame with all the data is produced by the code:

```
EBMT <- data.frame (ID=id,
                    born=rep(0,nsample),
                    start=rep(0,nsample),
                    end=end,
                    year=year,
                    agecl=agecl,
                    proph=proph,
                    match=match,i
                    dim=as.numeric(rep(1,length(id))),
                    ns=as.numeric(ns),
                    path=as.character(path),
                    f$d[,1:(max(ns)-1)])
```

Two attributes are added: the format of the event dates (days) and the transition matrix. Table 19 shows the first rows of the data frame.

The 2279 patients experienced 3255 transitions and 5534 episodes or spells. 785 recovered after the transplantation and 907 experienced the adverse event. 660 experience both a recovery and the adverse event. 370 experienced a relapse and 533 died before a relapse could occur. Patients who did not relapse or die before the end of the follow-up had censored observations (a total of 1376). 332 patients did not experience any event during the observation period. They are in the origin state T at the end of observation. Table 20 shows the numbers of transition. The table is produced by the function `Trans(EBMT)`.

Table 19 Biograph object. Selection of EBMT data

```
   ID born start  end      year agecl proph                 match idim ns path   Ev1 Ev2 Ev3
1   1    0     0  995 1995-1998 20-40    no no gender mismatch    1  2   TP  22.0  NA  NA
2   2    0     0  422 1995-1998 20-40    no no gender mismatch    1  4 TAZR  12.0  29 422
3   3    0     0 1264 1995-1998 20-40    no no gender mismatch    1  2   TA  27.0  NA  NA
4   4    0     0   84 1995-1998 20-40    no    gender mismatch    1  4 TAZR  42.0  50  84
5   5    0     0  114 1995-1998   >40    no    gender mismatch    1  3  TPR  22.0 114  NA
6   6    0     0 1427 1995-1998 20-40    no no gender mismatch    1  3  TAZ  27.0  33  NA
7   7    0     0  775 1995-1998   >40    no no gender mismatch    1  4 TAZD  28.5  29 775
8   8    0     0 1618 1995-1998 20-40    no no gender mismatch    1  2   TP  31.0  NA  NA
9   9    0     0 1111 1995-1998 20-40    no    gender mismatch    1  3  TAZ  29.0  87  NA
10 10    0     0  255 1995-1998 20-40    no no gender mismatch    1  2   TR 255.0  NA  NA
```

Table 20 Number of transitions and censoring

```
           Destination
Origin  T   P   A   Z   R   D Total Censored TOTAL
  T     0 785 907   0  95 160  1947      332  2279
  P     0   0   0 227 112  39   378      407   785
  A     0   0   0 433  56 197   686      221   907
  Z     0   0   0   0 107 137   244      416   660
  R     0   0   0   0   0   0     0        0     0
  D     0   0   0   0   0   0     0        0     0
  Total 0 785 907 660 370 533  3255     1376  4631
```

References

Blossfeld, H.P. and G. Rohwer (2002) Techniques of event history modeling. New approaches to causal analysis. Lawrence Erlbaum, Mahwah, New Jersey (2nd Edition).

Blossfeld, H.P., K. Golsh and G. Rohwer (2007) Event history analysis with Stata. Erlbaum, Mahwah, New Jersey.

De Graaf, A. and L. Steenhof (1999) Relatie en gezinsvorming van generaties 1945-1979: uitkomsten van het Onderzoek Gezinsvorming 1998 (Partnership and family formation of cohorts 1945-1979: results of the Netherlands Fertility and Family Survey 1998). *Maandstatiek Bevolking* 1999, December, pp. 21-36.

De Wreede, L.C., M. Fiocco and H. Putter (2011) mstate: An R package for the analysis of competing risks and multi-state models. *Journal of Statistical Software*, 38(7):1-30.

Fiocco M, Putter H, van Houwelingen HC (2008). Reduced-rank proportional hazards regression and simulation-based prediction for multi-state models. *Statistics in Medicine* **27**, 4340–4358.

International Institute for Population Sciences (IIPS) and Macro International. 2007. National Family Health Survey (NFHS-3), 2005–06: India: Volume I. Mumbai: IIPS.

Matsuo, H. and F. Willekens (2003) Event histories in the Netherlands Fertility and Family Survey 1998. A technical report. Research Report 03-1, Population Research Centre, University of Groningen. Available at http://www.rug.nl/prc/publications/researchreports/index

van Houwelingen HC, Putter H (2008). Dynamic predicting by landmarking as an alternative for multi-state modeling: an application to acute lymphoid leukemia data. *Lifetime Data Anal* **14**, 447–463.

Table 11 Biograph object. SHARELIFE data

```
         ID     born    start    end     country           IDc  cohort    sex eduf job1 children idim ns  path  Tr1  Tr2  Tr3  Tr4
12405 12405 1952.250 1952.250 2008.5       Spain ES-744755-01   1950+ female 1965 1972        2    1  2    HM 1975   NA   NA   NA
5639   5639 1937.917 1937.917 2008.5 Switzerland Cf-273420-01 1930-39 female 1956 1956       NA    1  3   HAM 1955 1979   NA   NA
14523 14523 1940.917 1940.917 2008.5      France FR-653700-01 1940-49 female 1960 1965        2    1  2    HM 1964   NA   NA   NA
9514   9514 1955.167 1955.167 2008.5     Denmark DK-380793-02   1950+ female 1975 1975        2    1  4  HACM 1973 1973 1976   NA
6247   6247 1922.917 1922.917 2008.5 Switzerland Cg-396802-01   <1930 female 1944 1944       NA    1  5 HMACM 1963 1967 1969 1988
19039 19039 1947.500 1947.500 2008.5       Italy IT-284033-01 1940-49 female 1967   NA        3    1  2    HM 1967   NA   NA   NA
22187 22187 1948.917 1948.917 2008.5 Netherlands NL-639208-02 1940-49   male 1963 1963        2    1  2    HM 1970   NA   NA   NA
23030 23030 1930.167 1930.167 2008.5      Poland PL-002013-01 1930-39 female 1940 1955        4    1  2    HM 1958   NA   NA   NA
12920 12920 1948.667 1948.667 2008.5      France FR-011657-02 1940-49 female 1965 1968        2    1  2    HA 1968   NA   NA   NA
811     811 1922.833 1922.833 2008.5     Austria AT-953411-01   <1930 female 1936   NA        1    1  4  HACM 1955 1955 1976   NA
```