How to create a *Biograph* object

Frans Willekens, 16 April 2013

A.1 Introduction

The purpose of this annex is to show how to create a *Biograph* object. The R code is included in the on-line documentation of the *Biograph* package (Version 2). In this note, eight data sets are used to illustrate the method. Two data sets are hypothetical data sets, the first carries information on three subjects and the second on 22 subjects. The third data set is the GLHS data .The fourth is the Netherlands Family and Fertility Survey 1989 (NLOG98). In the fifth example of how to prepare a *Biograph* object, data from the Survey of Health, Ageing and Retirement in Europe (SHARE) are used. The SHARE survey is modeled after the U.S. Health and Retirement Survey (HRS). Data from the National Family Health Survey of India are used in the sixth example. The NFHS is one of many Demographic and Health Surveys (DHS) organized in third-world countries and countries in transition. In the seventh example medical data are used. They are included in the *mstate* package for multistate modeling in R, developed by Putter and colleagues at Leiden University Medical Centre. The data cover 2279 leukemia patients who had a bone marrow transplant. Simulated life histories are used in the last example. The final example consists of simulated life histories.

The *Biograph* object carries information on subjects, transitions and observation period. The subject data consist of dates of birth and covariates, which may include time-varying covariates. In the GLHS data, marital status is a time-varying attribute. The age at marriage is included in the data. Information on transitions includes state sequences and ages at transition. The state sequence is the sequence of states occupied by a subject during the period of observation. The ages are ordered chronologically, with the age at the first transition displayed first, followed by the age at the second transition, etc.

A *Biograph* object is created in five steps. The first is the specification of the state space and the transitions between states. Transitions that are not possible or not relevant for the study are excluded. The transitions that are included are feasible and relevant. The second step is the selection of 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, data are stored in a data frame, two data attributes are attached. The first attribute indicates how dates are represented (e.g. calendar time, CMC or age) 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.

A.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 arrangements. Consider their living arrangements: living at the parental home (H), living alone (A), cohabiting (C) and married (M). The set of possible living arrangements constitutes the state space, which is denoted as {H, A, C, M}. The first person starts living independently on August 20, 2004 at the age of 18. It is her first transition, i.e. she leaves the parental home to live independently. She starts cohabitation on December 1, 2011 and is still cohabiting at the time of interview. The second person starts living independently in September 2011. The exact date is not known. He is still living independently at survey date. The third person starts living independently on August 10, 2006 and marries on March 16, 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 shown in Table A.1 A row carries information on an individual. A column has the date of entry in a given state.

Table A.1 Transition dates for three hypothetical individuals

	A	C	M
1	2004-08-20	2011-12-1	<NA>
2	2011-09-15	<NA>	<NA>
3	2006-08-10	<NA>	2012-03-16

The covariates are sex and level of education. The observation period differs between individuals. It starts at birth and ends at interview. The data are shown in Table A.2

Table A.2 Data on three hypothetical individuals

	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 code to produce a *Biograph* object is shown in `create.Simple1a.r` and `create.Simple1b.r`. The first step in the creation of a *Biograph* object is the specification of the state space. The state space is {H, A, C, M}. The second step is the selection of covariates. They are sex and education. The third stop is the specification of the observation period for each individual in the study. They are shown in Table A.2. 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 state sequences. 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 A.3 shows the object produced by the function Sequences.ind.0. The components f$d and f$path are included in the *Biograph* object.

Table A.3 Object produced by the *Biograph* function Sequences.ind.0

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

$d
      [,1]  [,2] [,3]
[1,] 12650 15309   NA
[2,] 15232    NA   NA
[3,] 13370 15415   NA

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

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 to assemble the data in a data frame and to add the date format and the parameters as attributes. The following code produces the *Biograph* object (Table A.4):

```
bio  <- data.frame (
                ID=id,
                born=born,
                start=start,
                end=interview,
                sex=sex,educ=educ,
                path=as.character(path),
                dates[,1:(max(ns)
                1)],stringsAsFactors=FALSE)
attr(bio,"format.date") <- "%Y-%m-%d"
attr (bio,"param") <- Parameters (bio)
```

Table A.4. *Biograph* object: hypothetical data A.

```
  ID       born      start        end sex   educ path        Tr1        Tr2
1  1 1986-04-05 1986-04-05 2019-05-09   F   High  HAC 2004-08-20 2011-12-01
2  2 1986-08-08 1986-08-08 2019-05-09   M Medium   HA 2011-09-15       <NA>
3  3 1986-11-28 1986-11-28 2019-05-09   F Medium  HAM 2006-08-10 2012-03-16
```

The data frame has different data types. The function str(bio) displays the data types (Table A.5):

Table A.5. *Biograph* object: data types.

```
'data.frame':	3 obs. of  9 variables:
 $ ID   : num  1 2 3
 $ born : chr  "1986-04-05" "1986-08-08" "1986-11-28"
 $ start: chr  "1986-04-05" "1986-08-08" "1986-11-28"
 $ end  : chr  "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
 $ path : chr  "HAC" "HA" "HAM"
 $ Tr1  : Date, format: "2004-08-20" "2011-09-15" "2006-08-10"
 $ Tr2  : Date, format: "2011-12-01" NA "2012-03-16"
 - attr(*, "format.date")= chr "%Y-%m-%d"
 - attr(*, "param")=List of 18
  ..$ nsample       : int 3
  ..$ numstates     : int 4
  ..$ namstates     : chr [1:4(1d)] "H" "A" "C" "M"
  ..$ absorbstates  : chr [1:2(1d)] "C" "M"
  ..$ iagelow       : num 0
  ..$ iagehigh      : num 34
  ..$ namage        : int  0 1 2 3 4 5 6 7 8 9 ...
  ..$ nage          : num 35
  ..$ maxtrans      : num 2
  ..$ ntrans        : int 3
  ..$ trans_possible: logi [1:4, 1:4] FALSE FALSE FALSE FALSE TRUE FALSE
...
  .. ..- attr(*, "dimnames")=List of 2
  .. .. ..$ Origin     : chr [1:4(1d)] "H" "A" "C" "M"
  .. .. ..$ Destination: chr [1:4(1d)] "H" "A" "C" "M"
  ..$ tmat          : 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" "C" "M"
  .. .. ..$ To  : chr [1:4(1d)] "H" "A" "C" "M"
  ..$ transitions   :'data.frame':	3 obs. of  6 variables:
  .. ..$ Trans: Factor w/ 3 levels "1","2","3": 1 2 3
  .. ..$ OR   : Factor w/ 2 levels "1","2": 1 2 2
  .. ..$ DES  : Factor w/ 3 levels "2","3","4": 1 2 3
  .. ..$ ORN  : Factor w/ 2 levels "A","H": 2 1 1
  .. ..$ DESN : Factor w/ 3 levels "A","C","M": 1 2 3
  .. ..$ ODN  : chr  "HA" "AC" "AM"
  ..$ nntrans       : num [1:4, 1:4] 0 0 0 0 3 0 0 0 0 1 ...
  .. ..- attr(*, "dimnames")=List of 2
  .. .. ..$ Origin     : chr [1:4(1d)] "H" "A" "C" "M"
  .. .. ..$ Destination: chr [1:4(1d)] "H" "A" "C" "M"
  ..$ locpat        : int 7
  ..$ ncovariates   : num 0
  ..$ covariates    : chr  "sex" "end"
  ..$ 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.

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 A.6.

Table A.6 *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
```

A.3 Hypothetical data B

Suppose we have information on a sample of 22 individuals. The state space consists of four fictitious states {H, A, B, C}. C is an absorbing state. Suppose that three transitions are possible: HA, AB and BC. Return transitions are not allowed. 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, six dates are given: the date of birth, the date at entry into observation, the date of interview and the dates of transitions between the states. 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 of that year, when he leaves H and enters A. On 23rd March, he experiences the second event, to state B. On 5th May he makes a transition to state C. He stays in that state until the end of observation on 25th May 2007. The data are shown in Table A.7.

The function `Sequences.ind.0` orders the dates chronologically and derives the state sequence. The components `f$d` and `f$path` 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”
attr(RS,"param") <- Parameters (RS)
```

The *Biograph* object is shown in Table A.8.

Table A.7 Hypothetical survey data: multiple transitions

```
   ID       Born      Start       Stop          A          B          C
   -  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
```

Table A.8 *Biograph* object: hypothetical data B

```
   ID       born      start        end cov path        Tr1        Tr2        Tr3
1   1 1991-05-14 2007-01-02 2007-05-25   X HABC 2007-02-11 2007-03-23 2007-05-05
2   2 1991-05-22 2007-01-17 2007-05-17   X   HA 2007-05-04       <NA>       <NA>
3   3 1991-12-27 2007-01-18 2007-05-10   X    H       <NA>       <NA>       <NA>
4   4 1991-01-01 2007-01-22 2007-05-13   X HABC 2007-02-28 2007-04-10 2007-05-10
5   5 1991-02-02 2007-02-10 2007-05-23   X   HA 2007-05-17       <NA>       <NA>
6   6 1991-06-08 2007-01-30 2007-05-15   X HABC 2007-02-12 2007-03-05 2007-04-17
7   7 1991-06-23 2007-04-04 2007-05-06   X    H       <NA>       <NA>       <NA>
8   8 1991-09-14 2007-04-29 2007-05-27   X    H       <NA>       <NA>       <NA>
9   9 1991-10-06 2007-05-18 2007-05-29   X    H       <NA>       <NA>       <NA>
10 10 1991-03-10 2007-05-20 2007-05-31   X    H       <NA>       <NA>       <NA>
11 11 1991-06-24 2007-05-15 2007-05-18   X    H       <NA>       <NA>       <NA>
12 12 1991-02-07 2007-02-05 2007-05-19   X HABC 2007-02-25 2007-04-01 2007-05-02
13 13 1991-06-01 2007-02-05 2007-05-10   X  HAB 2007-04-18 2007-04-30       <NA>
14 14 1991-06-14 2007-02-06 2007-05-28   X  HAB 2007-05-18 2007-05-20       <NA>
15 15 1991-10-07 2007-02-26 2007-05-22   X    H       <NA>       <NA>       <NA>
16 16 1991-04-07 2007-03-10 2007-05-25   X    H       <NA>       <NA>       <NA>
17 17 1991-10-21 2007-03-11 2007-05-12   X   HA 2007-05-08       <NA>       <NA>
18 18 1991-05-07 2007-03-28 2007-05-29   X    H       <NA>       <NA>       <NA>
19 19 1991-07-20 2007-03-15 2007-05-10   X HABC 2007-03-23 2007-04-08 2007-04-20
```

```
20 20 1991-09-05 2007-04-13 2007-05-20    X    H        <NA>       <NA>       <NA>
21 21 1991-09-15 2007-04-04 2007-05-11    X   HA 2007-05-09       <NA>       <NA>
22 22 1991-07-07 2007-04-25 2007-05-31    X HABC 2007-05-16 2007-05-20 2007-05-26
```

The data types in the data frame are shown in Table A.9. The code to produce the *Biograph* object is shown in `create.Simple2.r`.

Table A.9 *Biograph* object: data types.

```
'data.frame':      22 obs. of  9 variables:
 $ ID   : int  1 2 3 4 5 6 7 8 9 10 ...
 $ born : Date, format: "1991-12-30" "1991-01-09" "1991-01-19" ...
 $ 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" ...
 $ 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(*, "param")=List of 18
  ..$ nsample        : int 22
  ..$ numstates      : int 4
  ..$ namstates      : chr [1:4(1d)] "H" "A" "B" "C"
  ..$ absorbstates   : chr "C"
  ..$ iagelow        : num 15
  ..$ iagehigh       : num 17
  ..$ namage         : int  15 16 17
  ..$ nage           : num 3
  ..$ maxtrans       : num 3
  ..$ ntrans         : int 3
  ..$ trans_possible: logi [1:4, 1:4] FALSE FALSE FALSE FALSE TRUE
FALSE ...
  .. ..- attr(*, "dimnames")=List of 2
  .. .. ..$ Origin     : chr [1:4(1d)] "H" "A" "B" "C"
  .. .. ..$ Destination: chr [1:4(1d)] "H" "A" "B" "C"
  ..$ tmat           : 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"
  ..$ transitions   :'data.frame':  3 obs. of  6 variables:
  .. ..$ Trans: Factor w/ 3 levels "1","2","3": 1 2 3
  .. ..$ OR   : Factor w/ 3 levels "1","2","3": 1 2 3
  .. ..$ DES  : Factor w/ 3 levels "2","3","4": 1 2 3
  .. ..$ ORN  : Factor w/ 3 levels "A","B","H": 3 1 2
  .. ..$ DESN : Factor w/ 3 levels "A","B","C": 1 2 3
  .. ..$ ODN  : chr  "HA" "AB" "BC"
  ..$ nntrans        : num [1:4, 1:4] 0 0 0 0 12 0 0 0 0 8 ...
  .. ..- attr(*, "dimnames")=List of 2
  .. .. ..$ Origin     : chr [1:4(1d)] "H" "A" "B" "C"
  .. .. ..$ Destination: chr [1:4(1d)] "H" "A" "B" "C"
  ..$ locpat         : int 6
  ..$ ncovariates    : num -1
  ..$ covariates     : chr  "cov" "end" "start"
  ..$ format.date    : chr "%Y-%m-%d"
```

A.4 German Life History Survey

Throughout the documentation of the package, a single dataset is used to illustrate *Biograph*. The dataset is a subsample of the German Life History Survey (GLHS). The GLHS was organized in 1981-83 and provides information on the life histories of more than 5,000 men and women from three birth cohorts: 1929-31, 1939-41, and 1949-51. Blossfeld and Rohwer (2002) and Blossfeld et al. (2007) used a subsample of 201 respondents for training purposes. The 201 respondents experienced 600 job episodes. The data are used to illustrate hazard rate modelling of the job episodes with TDA (Transition Data Analysis) (2002 publication) and Stata (2007 publication). The same subsample of 201 respondents is used in this example. The example considers 201 *employment careers*, consisting of a total of 600 job spells and 382 episodes without a job. Dates of job entry and job exit are given in Century Month Code (CMC). Personal attributes are the date of birth and 5 covariates: sex, level of education, date of marriage, date of labour market entry and birth cohort. Marital status is a time-varying covariate. Education level is measured by 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.

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 programme `create.GLHS.r` converts the published data into a *Biograph* object. The programme is not part of the *Biograph* package but it is distributed with the package.

The published data file is an episode file. The filename is `rrdat`. The data are conveniently included in the *Biograph* package. The data object `rrdat.rda` can be retrieved by typing `data(rrdat)` after loading the *Biograph* package. The data file can also be downloaded from the designated website http://oldsite.soziologie-blossfeld.de/eha/tda/ using the following code, provided the computer is connected to the internet:

```
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)
```

A selection of the GLHS survey data is presented in A.10. 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. Observation starts at birth (TB) and ends at the date of interview (TI). A job episode is identified by a serial number (NOJ) and is characterized by 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).

Table A.10 GLHS input data for Blossfeld and Rohwer's TDA programme (rrdat)

(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

The *Biograph* object is prepared in five steps. The first is the specification of the state space and the possible transitions. The state space consists of two states: no job (N) and job (J). Everyone starts in state N. Transitions that are not feasible or not relevant for the study are excluded. The transitions that are included are feasible and relevant. The second step is the selection of covariates. The third step is the specification of the observation window for each subject. It requires the dates at start and end of observation. In the fourth step, the state sequence is determined and the dates at transition are recorded. In the fifth and final step, all data are stored in a data frame, two data attributes are attached to the data frame. The first attribute is the format of the dates and the second is a set of parameters that characterize the data. The parameters include the transition matrix, i.e. the matrix of possible and relevant transitions.

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. The authors 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.

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 the set of parameters:

```
attr(GLHS,"param") <- Parameters (GLHS)
```

The parameters include the matrix of feasible transitions, some packages require (`Parameters(GLHS)$tmat`).

Table A.11 shows a selection of the *Biograph* object. In the *Biograph* object a record contains the following variables (columns):

- `ID`: identification number of respondent. ID is a numeric value. The values do not need to be sequential, but they need to be numeric. Character variables are not allowed.
- `born`: date of birth of respondent. The date may be in CMC or another date format. The date format is one of the attributes of the *Biograph* object.
- `start`: onset of observation
- `end`: end of observation
- Four covariates: sex (`sex`), highest educational attainment before entry into the labour market (`edu`), date (CMC) of marriage (`marriage`) and date (CMC) of entry into the labour market
- `path`: sequence of states occupied during the observation period. The variable `path` must be a character variable. Each state is represented by a single character.
- `Tr*:` transition dates in CMC. The maximum number of transitions is determined by the data. In this subsample, it is 12.

The variable `path` (for lifepath) is a character variable representing the sequence of states occupied during the observation period. The `path` should be a character variable, otherwise *Biograph* gives an error message and stops. You should check that the variables are of the required type, using the `str(GLHS)` command. The variables `pres` and `NOJ` are omitted since these variables are associated with job episodes and not with persons.

Table A.11 *Biograph* object: GLHS data

```
   ID born start end    sex edu marriage LMentry  cohort       path Tr1 Tr2 Tr3 Tr4 Tr5 Tr6 Tr7
1   1  351   351 983   Male  17      679     555 1929-31         NJ 555  NA  NA  NA  NA  NA  NA
2   2  357   357 983 Female  10      762     593 1929-31      NJJJN 593 639 673 893  NA  NA  NA
3   3  473   473 983 Female  11      870     688 1939-41    NJJJJJN 688 700 730 742 817 829  NA
4   4  604   604 983 Female  13      872     872 1949-51        NJN 872 927  NA  NA  NA  NA  NA
5   5  377   377 983   Male  11      701     583 1929-31       NJJJ 583 651 788  NA  NA  NA  NA
6   6  492   492 983   Male  11      781     691 1939-41   NJNJNJNJ 691 717 728 754 771 847 859
7   7  476   476 983 Female   9      748     652 1939-41     NJJJJN 652 705 730 736 751  NA  NA
8   8  609   609 983   Male  11      881     838 1949-51       NJJJ 838 844 892  NA  NA  NA  NA
9   9  377   377 983   Male  12      690     591 1929-31      NJJJJ 591 602 634 643  NA  NA  NA
10 10  382   382 983   Male  11      824     580 1929-31      NJJNJ 580 701 843 862  NA  NA  NA
```

The length of the string `path` gives the number of states an individual occupies during the period of observation. The number of states individual i occupies is `nchar(GLHS$path[i])`. Who is the respondent occupying the largest number of states in the observation period? The answer is obtained using the following command:

```
    GLHS[nchar(GLHS$path)==max(nchar(GLHS$path)),]
```

A.5 Netherlands Family and Fertility Survey

Between February and May 1998 Statistics Netherlands (CBS) conducted the Netherlands Family and Fertility Survey. Data were collected on 5,450 women and 4,717 men in the Netherlands, born in the period 1945-79. They were 18 to 52 years at time of survey. The sample frame consisted of the Municipal Population Administration (Gemeentelijke Bevolkingsadministratie or GBA). The GBA is the main source of statistical information on the population in the Netherlands. The random sample survey was done in two steps. In the first step 262 municipalities were selected from 572 municipalities. GBA data of the selected municipalities were then used to randomly select 14 thousand addresses and subsequently men and women born in the period of 1945-1979. For details on the sampling, see de Graaf and Steenhof, 1999, p.36). Eventually, 5450 women and 4717 men were interviewed using structured questionnaires.

DANS (Data Archiving and Networked Services) distributes the survey data for public use (https://easy.dans.knaw.nl/; search for *gezinsvorming*). The data are distributed in two SPSS files. The file BOAV98.SAV (or *.POR) contains the data for females and the file BOAM98.SAV (or *.POR) contains the data for males. In this chapter, data on females are used.

The Netherlands Family and Fertility Survey provides extensive information on marital status, living arrangements, partnership and fertility. The information is collected retrospectively and covers the period from birth to survey date. 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 raw data need considerable processing to be useful for *Biograph*. First, the public use file does not include the survey month. Although we know that the survey took place in the period 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 survey month is estimated from the age at survey, the month of birth of the respondent and the months in which transitions occur. No transition may occur after the survey date. The estimation procedure includes a random number generation to allocate the survey date to one of several plausible months, taking into account that transitions reported by the respondent could not have taken place after the survey date.

Second, the public use data file is not well suited for life history data analysis. The focus of the questionnaire is 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). In some cases imputation was necessary. The emphasis on the sequence and timing of events did reveal several inconsistencies in the data. Some sequences of events are not possible (e.g. second child is born before first child) or are not plausible (e.g. remarriage before a divorce). 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) is limited to the data for females. Starting from the public use file BOAV98.SAV, inconsistencies are removed and an event history data file prepared in10 steps. Each step is documented in an SPSS syntax file. The report by Matsuo and Willekens and the SPSS syntax files 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 name of the SPSS file with the event histories is NLOG98_F_CMC.sav. The syntax file also creates a text file in *Biograph* format to be used as input in *Biograph*. The name of that data file is NLOG98cov.DAT. For the illustrations in this chapter, a subsample of 500 women was selected from the 5450 women in the NLOG98 sample. That file is included in the *Biograph* package under the name `NLOG98.Rdata`. The command `data(NLOG98)` loads the data set. For convenience, the data object is named renamed to `OG`.

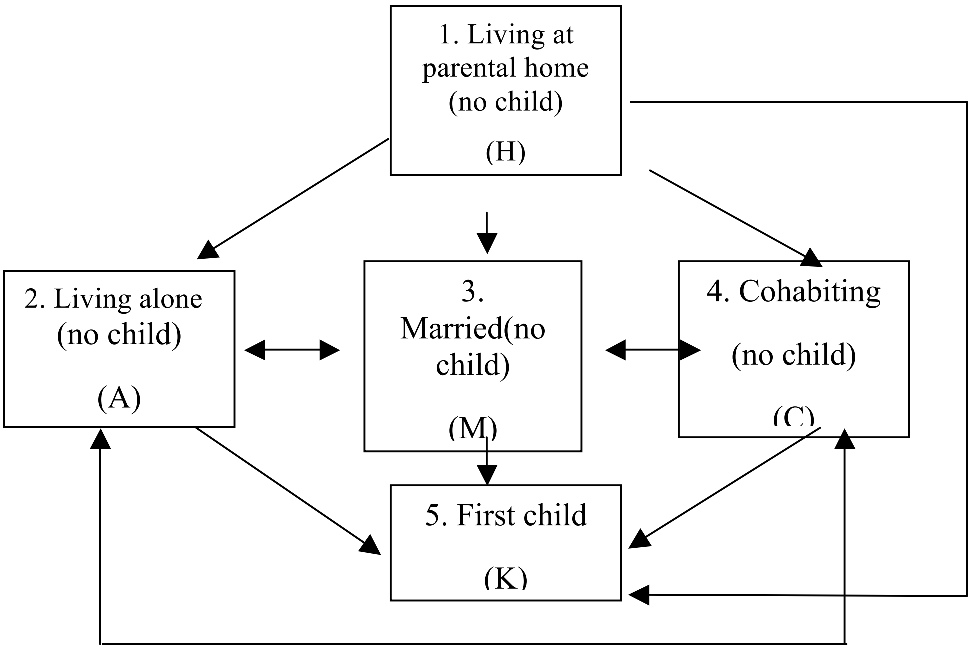
A *Biograph* object is created in five steps. The steps are implemented in the programme `create.NLOG98.r`, which is distributed with the *Biograph* package (see Documentation folder of the package source `Biograph_2.0.2.tar.gz` or later version). The first step is the specification of the state space and the possible transitions. The second step is the selection of covariates. The third step is the specification of the observation window for each subject. In the fourth step, the state sequence is determined and the dates at transition are recorded. In the fifth and final step, all data are stored in a data frame, two data attributes are attached to the data frame. The state space describes the pathways to the first child, i.e. the set of states a woman may occupy before the first child is born. Figure 8.1 presents the state space and the associated transitions. The 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 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 domains of life are combined into a single state space. Some combinations of states are excluded (e.g. cohabitating at the parental home, married while living at the parental home). The primary states of interest are:

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

The specification of the state space determines the sequence of states and events that can be studied. In this example, a married woman may start cohabitation upon marriage dissolution. She may start living alone instead but she may not move back to the parental home. Some living arrangements, such as Living Apart Together (LAT) (commuting marriage), are not considered in the state space and can therefore not be studied. To include that arrangement a distinction must be made between partnership status (union status) and residence status, and the timing of the transitions between the states should be known. The focus on pathways to first birth implies that the transitions that occur after the birth of a child are not considered in the analysis. The birth of a child implies entry into an absorbing state.

The NLOG98 reveals some uncommon living arrangements. For instance, some married women do not live with their husband; some live alone and some live with another partner. These living arrangements are not considered in this report since we lack information and the state space is too restrictive. To capture these living arrangements, the state space would need to be extended.

Figure 8.1. Schematic representation of pathways to first child


1. Living at parental home (no child) (H)
2. Living alone (no child) (A)
3. Married(no child) (M)
4. Cohabiting (no child) (C)
5. First child (K)

The state space has five states with the names "H" (living at the parental home), "A" (living alone), "C" (cohabiting), "M" (married) and "K" (with at least one child). The number of possible transitions is 16: HA, HM, HC, HK, AM, AC, AK, CA, CM, CK, MA, MC, MK, AA, CC and MM. Cohabiting or married respondents may change partner without a period of independent living, resulting in transitions CC and MM. That transition is maintained in the data set, but it is disregarded because it is a transition to the same state. In the subsample, 2 women reported a CC transition. The feasible transitions are identified in the transition matrix:

```
                 Destination
Origin     H     A     C     M     K
     H FALSE  TRUE  TRUE  TRUE  TRUE
     A FALSE FALSE  TRUE  TRUE  TRUE
     C FALSE  TRUE  TRUE  TRUE  TRUE
     M FALSE  TRUE  TRUE FALSE  TRUE
     K FALSE FALSE FALSE FALSE FALSE
```

Most packages for multistate analysis with R construct and use a transition matrix. In *Biograph*, The transition matrix is contained in the object `attr (OG,"param")$trans_possible` (see below). The calendar dates of transitions are expressed in Century Month Code (CMC).

To determine the dates at transition, the following date variables are extracted from the data file NLO98_F_CMC:

	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

The second step is the selection of covariates. Two covariates are selected and included in the *Biograph* object: religious denomination (kerk) and level of education (educ). The first covariate is religion (labeled KERKGEZ in the original data distributed by Statistics Netherlands). The following categories are distinguished, with the original code in brackets and the number of respondents n in the original sample of 5450 respondents:

1. No religion [1] (n = 2395)
2. Roman Catholic [2] (n = 1677)
3. Protestant [3, 4, 5 and 6] (n = 1014)
4. Other religion [7, 8, 9 and 10] (n = 357)

NA Missing data [98, 99] (n = 7)

The second covariate is the highest completed education. In the original data set, the covariate is called OPL_HB. The following categories are distinguished, with the original codes in brackets and the number of respondents n:

1. Primary [2] (n = 363)
2. Secondary lower [3] (n = 1250)
3. Secondary higher [4] (n = 2489)
4. First step high [5] (n = 869)
5. Second step high [6] (n = 238)
6. Third step high [7] (n = 20)

NA Missing data [9] (n = 221)

In addition, two birth cohorts are derived from the dates at birth. The first cohort is born before 1960 and the second cohort is born in 1960 or later.

The third step is the specification of the observation window for each subject. The life history is recorded retrospectively starting at birth and ending at interview date. The interview date is given in CMC and the assumption is made that interview is at the end of the month, estimated using the procedure described above. Since *Biograph* assumes that events, including censoring, occur at the beginning of a month, a one is added to the interview month.

In the fourth step, the state sequence is determined and the dates at transition are recorded. The *Biograph* function `Sequences.ind.0` is used. The function orders dates chronologically and determines the state sequence. The output is an object with four components. Two components are included in the *Biograph* object: (1) the character string denoting the state sequence (`Sequences.ind.0$path`) and (2) the sequence of the CMCs at transition (`Sequences.ind.0$d`).

In the fifth and final step, all data are stored in a data frame, two data attributes are attached to the data frame (the '`format.date`' attribute and the '`param`' attribute).

A selection of the subsample of 500 respondents is shown in Table A.12. The variables `ID`, `born`, `start`, `end` and `Tr*` are numeric. The variable `path` is a character variable and the covariates are factors.

Table A.12 *Biograph* object: selection of NLOG98 data.

```
   ID born start  end           kerk educ cohort   path  Tr1  Tr2  Tr3  Tr4  Tr5
2   2  630   630 1184    no religion    5  <1960    HAC  966 1002   NA   NA   NA
8   8  707   707 1180 Roman Catholic   NA  <1960   HCMK  894  906  910   NA   NA
24 24  813   813 1179 Roman Catholic    4  1960+    HCK 1004 1040   NA   NA   NA
28 28  673   673 1180    no religion    2  <1960   HMCK  939  990 1066   NA   NA
34 34  789   789 1179    no religion    6  1960+   HACA 1016 1105 1150   NA   NA
43 43  609   609 1179     Protestant   NA  <1960     HK  840   NA   NA   NA   NA
52 52  895   895 1182    no religion    5  1960+     HA 1118   NA   NA   NA   NA
82 82  689   689 1181    no religion    6  <1960   HACM  973 1003 1013   NA   NA
96 96  721   721 1182    no religion    4  <1960 HACACK 1034 1038 1111 1128 1140
99 99  862   862 1181    no religion    4  1960+    HAC 1089 1105   NA   NA   NA
```

A.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 prepare the *Biograph* object is shown in `create.SHARElife.r`. 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 SHARELIFE data are used to investigate how living arrangements change with age. The observation period is from birth to survey date. 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 ended
- Year in which respondent start first job
- Birth cohort: four birth cohorts: <1930, 1930-39, 1940-49, 1950+

The variables that are extracted from the raw data are identification number, date of birth, date of interview, dates of transition and selected covariates:

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$sl.rp003.n	Year in which cohabitation NOT related to marriage started (n = 11 to 18)
•	d.rp$sl.rp012.n	Year in which cohabitation NOT related to marriage 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:

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

The variables are:

`bb`	date of birth in decimal year
`end`	date of interview in decimal year
`d.st$country`	country
`Idc`	identification number used in SHARELIFE (character variable)
`bcohort`	birth cohort
`sex2`	sex (“male” and “female”)
`edu.f`	year in which full-time education is finished
`job.1.start`	year of entry in labour market
`nchildren`	number of children

Two attributes are added to the data file: the format of the transition dates (year) and the set of parameters. For one respondent the date of birth is missing; he is removed from the data.

Table A.13 shows a selection of rows of the SHARELIFE data in the *Biograph* format.

A.6 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 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. Therefore a one is added to variable v008.

The raw data present 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 A.14 shows a selection of rows.

Table A.14 *Biograph* object: NFHS-AP

```
   ID born start  end     COH EDU WEAL U_R CEB     path  Tr1  Tr2  Tr3  Tr4  Tr5  Tr6
1   1  709   709 1274   <1970   0    2   2   4  HMabcdS  936  937  964 1006 1045 1045
2   2  997   997 1274  >=1980   1    2   2   2    HMabS 1200 1210 1238 1238   NA   NA
3   3 1033  1033 1276  >=1980   0    2   2   1      HMa 1172 1197   NA   NA   NA   NA
4   4 1009  1009 1274  >=1980   0    3   2   2    HMabS 1193 1202 1221 1221   NA   NA
5   5  973   973 1274  >=1980   2    3   2   2    HMabS 1169 1200 1211 1211   NA   NA
6   6  733   733 1274   <1970   0    4   2   3   HMabcS  919  949  997 1040 1046   NA
7   7  985   985 1274  >=1980   2    4   2   2    HMabS 1241 1250 1262 1263   NA   NA
8   8 1011  1011 1274  >=1980   0    3   2   1      HMa 1205 1238   NA   NA   NA   NA
```

The code is shown on `create.NFHS.r`.

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

a. Introduction

The EBMT data are included in the *mstate* package, developed by Putter and colleagues (see de Wreede et al., 2011).

The hematopoietic stem cells in bone marrow in large bones 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. Engraftment is the process of transplanted stem cells reproducing new 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 data, in a file names `ebmt4` included in the *mstate* package, 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:

i. Acute graft versus host disease (AGvHD). AGvHD is a GvHD of grade 2 or higher, appearing before 100 days post-transplant.
ii. Platelet recovery. 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. Platelet recovery is the recovery of platelet count**.**
iii. Relapse and death.

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 two intermediate events (recovery and an adverse event) is used to update the prognoses of the patients. An example of an adverse event is an 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. Instead of Recovery, Engraftment or platelet recovery can be included. The multistate model considers six states (with the multi-character state labels used in *mstate* and the single-character state labels used in *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 experiences 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 dies after 579 days. The last event is not relevant to the model because the patient has already reached an absorbing state.

Putter et al. make a few adjustments of the data for a multi-state analysis. Since the model does not allow patients to enter two states at the same time, a patient who experiences relapse and death on the same day is assumed to have entered the absorbing state of relapse rather than death, because the patients experience relapse before death. Patients who experience the adverse event and recovery on the same day are assumed to experience 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. The data include four covariates: year at transplantation, age at transplantation, donor-recipient gender match and prophylaxis.

d. Preparation of *Biograph* object

The preparation of a *Biograph* object involves the five steps listed in previous sections of the annex. The code is shown in `create.ebmt.r`. The state space includes the six states shown above: {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 A.15 shows the first rows of the data frame. The maximum number of transitions patients experience is 3.

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

Table A.15 Data frame with event dates in days since transplantation. EBMT.

```
    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 function `Sequences.ind.0` arranges event dates chronologically and determines the state sequence:

```
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,
                    path=as.character(path),
                    f$d[,1:(max(ns)-1)])
```

Two attributes are added: the format of the event dates (days) and the set of parameters. Table A.16 shows the first rows of the data frame.

Table A.16 *Biograph* object: EBMT data

```
   ID born start  end       year agecl proph                match path   Tr1 Tr2 Tr3
1   1    0     0  995 1995-1998 20-40    no no gender mismatch   TP  22.0  NA  NA
2   2    0     0  422 1995-1998 20-40    no no gender mismatch TAZR  12.0  29 422
3   3    0     0 1264 1995-1998 20-40    no no gender mismatch   TA  27.0  NA  NA
4   4    0     0   84 1995-1998 20-40    no    gender mismatch TAZR  42.0  50  84
5   5    0     0  114 1995-1998   >40    no    gender mismatch  TPR  22.0 114  NA
6   6    0     0 1427 1995-1998 20-40    no no gender mismatch  TAZ  27.0  33  NA
7   7    0     0  775 1995-1998   >40    no no gender mismatch TAZD  28.5  29 775
8   8    0     0 1618 1995-1998 20-40    no no gender mismatch   TP  31.0  NA  NA
9   9    0     0 1111 1995-1998 20-40    no    gender mismatch  TAZ  29.0  87  NA
10 10    0     0  255 1995-1998 20-40    no no gender mismatch   TR 255.0  NA  NA
```

A.9 Simulated life histories

Life histories of 200 individuals are simulated from age 20 to age 40. Individuals can occupy one of three states, labeled A, B and C. A state is selected randomly for each of the 200 individuals. Transition rates are constant between ages 20 and 40. The transition matrix **M**, with origin in column and destination in row, is:

$$\mathbf{M} = \begin{bmatrix} 0.15 & -0.07 & -0.02 \\ -0.10 & 0.10 & -0.05 \\ -0.05 & -0.03 & 0.07 \end{bmatrix}$$

The generate the life history of a single individual, the function `sim.msm` of the *msm* package is used. The function simulates an individual trajectory from a continuous-time Markov model (Jackson, 2013). The function requires the transition matrix in a different format than shown above and used in this book. The row variable should indicate origin and the column variable destination. The off-diagonal elements should be transition rates rather than minus transition rates. The required format is produced by –t(**M**), where t() denotes transpose. The function also requires that a numeric value rather than a character denotes a state. A is replaced by 1, B by 2 and C by 3. The simulation is an application of dynamic microsimulation in continuous time. For background information, see e.g. Willekens (2009).

The following code generates a trajectory for an individual starting in state A (state 1) at age 20:

```
bio <- sim.msm (-t(M),mintime=20,maxtime=40,start=1)
```

The object produced by `sim.msm` has three components. The first is the state sequence. The second provides information on the start of the observation window, transition times, and end of the observation window. The third is the matrix of transition rates. The result is:

```
$states
[1] 1 3 1 2 2
$times
[1] 20.00000 21.57682 38.39406 39.17881 40.00000
$qmatrix
      destination
origin     A     B     C
     A -0.15  0.10  0.05
     B  0.07 -0.10  0.03
     C  0.02  0.05 -0.07
```

The subject starts in A, moves to C, back to A and continues to B. At the end of the observation period, the individual is in B. The first transition is at age 21.58, the second at 38.39 and the third at 39.18. The character string showing the state sequence is ACAB.

The code `Create.simul.r` generates trajectories for 200 individuals. Table A.17 shows the data for the first 10 individuals.

Table A.17 *Biograph* object: simulated life histories.

```
ID born start end cov1   path   Tr1   Tr2   Tr3   Tr4   Tr5
 1    0    20  40    X     AB 28.14    NA    NA    NA    NA
 3    0    20  40    X     AB 23.69    NA    NA    NA    NA
 4    0    20  40    X     BA 24.56    NA    NA    NA    NA
 5    0    20  40    X   ACBC 20.73 27.79 37.81    NA    NA
 6    0    20  40    X BACABC 30.15 33.33 34.29 35.87 38.00
 7    0    20  40    X    CAC 29.44 34.39    NA    NA    NA
 8    0    20  40    X   CBCB 34.32 35.56 39.44    NA    NA
 9    0    20  40    X      B    NA    NA    NA    NA    NA
10    0    20  40    X  ABABA 22.80 25.29 26.73 39.37    NA
```