

Analysis plan, data requirements and analysis of HH events in DK

SDC

<http://bendixcarstensen.com/SDC/KZIA/HH-DK>

September 2016

Version 6

Compiled Monday 3rd October, 2016, 13:40
from: /home/bendix/sdc/coll/kzia/r/HH-DK.tex

Bendix Carstensen Steno Diabetes Center, Gentofte, Denmark
& Department of Biostatistics, University of Copenhagen
bxc@steno.dk
<http://BendixCarstensen.com>

Contents

1	Analysis plan for HH	1
1.1	Age, date and duration at entry	2
1.2	Reporting estimates	2
2	Data requirements	3
3	Reading and grooming data	4
4	Analysis of HH rates	15
5	Overall rates by calendar time	17
6	Analysis by HH status	20
6.1	Clinical variables	21
6.1.0.1	Persons with at lest one event	23
7	Modeling HH rates	24
7.0.1	Occurrence rates for 1st HH	30
8	Sensitivity modeling	31
8.1	Including albumin in the model	32
8.2	Categorizing variables	33
9	Overview of baseline data	34

1 Analysis plan for HH

The core of the analysis is to show how the HH-rates change by calendar time, while taking sex, age, diabetes duration and clinical variables, notably previous HH occurrence into account.

Thus a *single* statistical model describing the occurrence of HH in the period 2006-01-01 through 2012-12-31 must be set up, with effects of the clinical and demographic variables.

For simplicity the model will have (linear) effects of the prominent clinical variables (measured at entry): BMI, HbA_{1c}, albumin level (log₁₀-transformed), systolic blood pressure, as well as the categorical variables smoking, BP-treatment and Lipid-treatment. Finally we should include the number of previous HH events recorded, as well as the time since last HH event.

This is essentially a multistate model as shown in figure 1 (it can of course always be discussed how many instances of HH event we should count, in the figure we chose 5 purely for convenience):

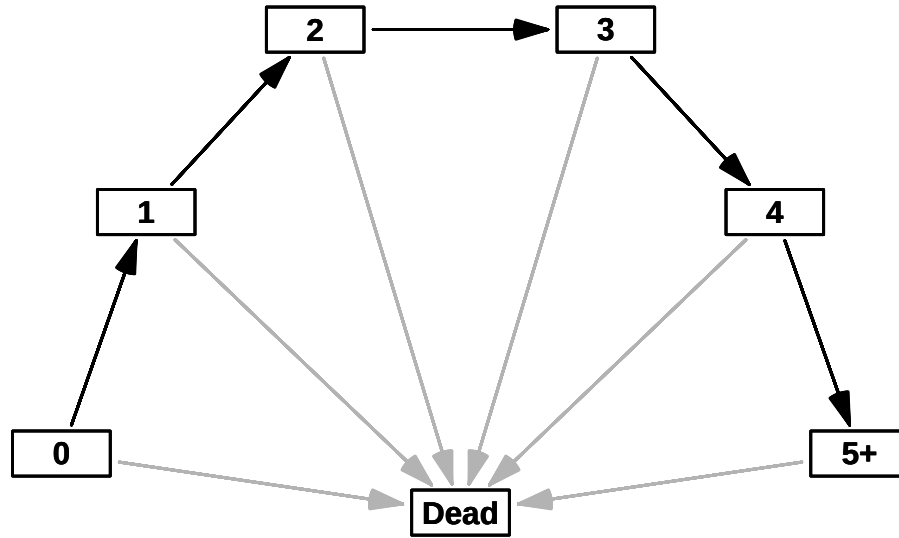


Figure 1: *The possible states of HH (shown as the number of previous HH) recorded. The black arrows represent the incidence rates of HH given a certain number of previously recorded HH. These are the rates of interest in this study. The gray rates are mortality rates that are not analysed here.*

We set up a model for the HH rates, λ , with terms for the clinical variables recorded at entry in $X\beta$, and with timescales age (a), duration of DM (d), calendar time (t) and time since last HH (h) (which is HH event number H):

$$\log(\lambda) = X\beta + f(a) + g(d) + k(t) + \ell_H(h)$$

where f , g and k are common smooth functions of the timescales — these would be reported as HRs relative to some chosen reference points on the yime scales. The ℓ_H

are *different* smooth functions of time since latest HH, one for each number of previous HH, $H=1,2,3,4$, interpretable as the incidence HR of HH relative to those persons without. The model thus assumes that the only thing that differs between the transition rates by history of HH is the effect of time since last HH. A simpler model would assume that these functions were parallel, differing by a constant only:

$$\ell_H(h) = \delta_H \ell_0(h)$$

where the δ s are then interpretable as the RR of HH between persons with different number of previous HHs — a type of proportional hazards assumption (no interaction between number of HH and time since last HH):

$$\ell_H(h) = \delta_H$$

An even simpler assumption would be to assume that the rates of HH did not depend on time since last HH:

1.1 Age, date and duration at entry

Note that the model does *not* (directly) include the differences between the timescales:

$a - d$:	age at diagnosis of DM
$a - h$:	age at last HH
$d - h$:	diabetes duration at last HH
$t - d$:	date of DM diagnosis
$t - h$:	date of last HH

Since each of the time scale effects are assumed non-linear (but including the linear component) the linear effect of these variables are accommodated in the model, but possible non-linear effects are not included.

Also note that there is no way in the proposed model to disentangle the linear effects of current age (a) and age at diagnosis or age at last HH.

The way we have chosen to parametrize the model, the age effect is the so-called *cross-sectional* age-effect, that is how incidence rates vary by age for *fixed* values of DM duration, time since last HH and calendar time. The age-effects are thus not interpretable as the change in rates a group of people see as they age.

1.2 Reporting estimates

The reporting of the estimated effects will be both as separate plots of the smooth functions as RR relative to some suitable reference point; age $a = 50$, diabetes duration $d = 25$, and calendar time $t = 2010$, say. Note that no reference point is needed for the time since last HH because this variable is undefined for persons without previous HH.

To illustrate the interplay between the timescales we shall also show the incidence rate by calendar time for persons diagnosed at ages (say) 20, 30, 40 at 1 January 2005, 2007, 2009 and 2011, and followed through 2012-12-31. This will produce 12 curves, 3 starting in each year, corresponding to the different ages at diagnosis. The time since last HH should be fixed in these plots, and they should be shown for persons without previous HH.

The same plot should be shown for persons diagnosed in the same ages but 10, resp. 20 years earlier. These will also have 12 curves but all stretching over the entire period.

All these plots would of course be for specific values of the clinical variables to be fixed. The analysis should show the effect of previous events of HH by including the number of HH and the time since last HH.

Thus all transitions rates shown in black in figure 1 will be considered, and they will be assumed to depend on the clinical variables in the same way.

2 Data requirements

In order to do this analysis a dataset with one record per person is required, with the following variables (clinical variable refer to date of entry to the study). Note that only the date of the last HH before entry to the study need be given — from date of entry can then be inferred the state (no. of previous HH) the person starts in.

<code>id</code>	person id
<code>dob</code>	date of birth
<code>dodm</code>	date of diabetes
<code>doe</code>	date of entry to the study
<code>dox</code>	date of exit from the study
<code>doh1</code>	date of 1 st HH
<code>doh2</code>	date of 2 nd HH
	... etc.
<code>dod</code>	date of death
<code>sex</code>	sex (m/f)
<code>hba</code>	HbA _{1c} (mmol/mol) at <code>doe</code>
<code>bmi</code>	body mass index (kg/m ²) at <code>doe</code>
<code>alb</code>	albumin level (mg/l/day) at <code>doe</code>
<code>sbp</code>	systolic blood pressure (mmHg) at <code>doe</code>
<code>aht</code>	anti hypertensive treatment (y/n) at <code>doe</code>
<code>llt</code>	lipid lowering treatment (y/n) at <code>doe</code>
<code>smk</code>	smoking status (never/ex/occ/curr) at <code>doe</code>

The dates must obey the obvious requirements of internal ordering. The variables in the dataset must have precisely the names given above, and no other variables should be in the dataset.

3 Reading and grooming data

Data are however made available in a slightly different form; for persons with at least one HH, one record per HH, in a variable `doh` (and all other variables identical across records within each person), and for persons without any HH one record with `doh` coded as missing:

```
> options( width=90 )
> library( Epi )
> hh <- read.csv( "../data/t1new.csv", as.is=TRUE )
> # Reorder variables sensibly
> hh <- hh[,c(2:1,4:9,3,10:ncol(hh))]
```

We can then provide an overview of the original dataset, which we shall use as basis for construction of a Lexis representation of follow-up through states of increasing number of HHs:

```
> dim( hh )
[1] 28634    15
> names( hh )
 [1] "id"   "dob"   "dodm"  "doe"   "dox"   "doh"   "dod"   "sex"   "bmi"   "hba"   "alb"   "sbp"
[13] "aht"   "llt"   "smk"
> hh <- transform( hh, sex = factor( sex, levels=1:2, labels=c("M","F") ),
+                 aht = factor( aht, levels=1:2, labels=c("N","Y") ),
+                 llt = factor( ll, levels=1:2, labels=c("N","Y") ),
+                 smk = Relevel( factor(smk), "Never smoker" ),
+                 gbmi = cut( bmi, breaks=c(-Inf,20,25,30,Inf), right=FALSE ),
+                 galb = cut( alb, breaks=c(-Inf,30,300,Inf), right=FALSE ) )
> levels( hh$gbmi ) <- c("Under-wt","Normal","Over-wt","Obese")
> levels( hh$galb ) <- c("Normo","Micro","Macro")
> dvar <- grep( "do", names(hh) )
> for( i in dvar ) hh[,i] <- as.Date( hh[,i], format="%Y-%m-%d" )
> hh <- cal.yr( hh )
> or <- with( hh, order(id,doh) )
> hh <- hh[or,]
> str( hh )

'data.frame':      28634 obs. of  17 variables:
 $ id   : int  1 3 4 7 8 9 10 11 11 11 ...
 $ dob  : num  1954 1972 1968 1935 1951 ...
 $ dodm : num  2003 1980 2002 1969 1996 ...
 $ doe  : num  2007 2006 2006 2009 2008 ...
 $ dox  : num  2012 2013 2012 2012 2010 ...
 $ doh  : num  NA NA NA NA NA NA ...
 $ dod  : num  NA NA NA NA NA NA NA NA NA ...
 $ sex  : Factor w/ 2 levels "M","F": 1 2 2 1 1 2 1 1 1 1 ...
 $ bmi  : num  26 26.1 25.6 23.5 24.8 ...
 $ hba  : num  39.9 76 61.7 45.4 42.1 ...
 $ alb  : num  NA NA NA 0.1 NA ...
 $ sbp  : int  134 160 144 135 135 145 142 115 115 115 ...
 $ aht  : Factor w/ 2 levels "N","Y": 1 1 1 2 2 1 1 2 2 2 ...
 $ llt  : Factor w/ 2 levels "N","Y": 1 1 1 2 2 1 1 1 1 1 ...
 $ smk  : Factor w/ 5 levels "Never smoker",...: 1 2 2 3 2 2 2 4 4 4 ...
 $ gbmi : Factor w/ 4 levels "Under-wt","Normal",...: 3 3 3 2 2 4 2 2 2 2 ...
 $ galb : Factor w/ 3 levels "Normo","Micro",...: NA NA NA 1 NA NA 1 1 1 1 ...

> summary(hh)

      id      dob      dodm      doe      dox
Min.   :      1   Min.   :1913   Min.   :1933   Min.   :2005   Min.   :2007
1st Qu.: 5654    1st Qu.:1949    1st Qu.:1976    1st Qu.:2007    1st Qu.:2012
Median :11596    Median :1961    Median :1988    Median :2007    Median :2012
```

```

Mean      :11560   Mean      :1961   Mean      :1986   Mean      :2008   Mean      :2012
3rd Qu.   :17324   3rd Qu.   :1973   3rd Qu.   :1998   3rd Qu.   :2009   3rd Qu.   :2013
Max.      :23268   Max.      :1995   Max.      :2012   Max.      :2013   Max.      :2013

      doh      dod      sex      bmi      hba
Min.    :1985   Min.    :2007   M:17073   Min.    :14.45   Min.    : 22.40
1st Qu. :2000   1st Qu. :2009   F:11561   1st Qu. :22.28   1st Qu. : 56.00
Median  :2005   Median  :2010                Median :24.38   Median : 65.03
Mean    :2005   Mean    :2010                Mean   :25.02   Mean   : 66.63
3rd Qu. :2009   3rd Qu. :2011                3rd Qu.:27.15   3rd Qu.: 74.86
Max.    :2014   Max.    :2012                Max.   :62.87   Max.   :249.73
NA's    :12440   NA's    :26927

      alb      sbp      aht      ll      smk
Min.    : 0.00   Min.    : 0.0   N:17244   N:18857   Never smoker :10211
1st Qu. : 0.50   1st Qu. :120.0 Y:11390   Y: 9777   Daily smoker  : 9077
Median  : 1.13   Median :130.0                Ex smoker     : 3830
Mean    : 11.11   Mean    :132.7                No information : 4591
3rd Qu. : 2.90   3rd Qu. :142.0                Occasional smoker: 925
Max.    :1072.03 Max.    :250.0
NA's    :11386   NA's    :1048

      gbmi      galb
Under-wt: 2185   Normo:16223
Normal   :13904   Micro: 869
Over-wt  : 9485   Macro: 156
Obese    : 3060   NA's :11386

```

```
> head(hh)
```

```

  id  dob  dodm  doe  dox doh dod sex  bmi  hba alb sbp aht ll      t
1  1 1954.419 2002.999 2007.158 2012.420 NA NA  M 25.96305 39.89085 NA 134  N  N
2  3 1971.676 1979.870 2006.200 2012.768 NA NA  F 26.14269 75.95655 NA 160  N  N
3  4 1968.335 2002.038 2006.285 2012.335 NA NA  F 25.59374 61.74885 NA 144  N  N
4  7 1935.027 1969.001 2009.028 2012.026 NA NA  M 23.45856 45.35535 0.1 135  Y  Y
5  8 1950.646 1995.999 2007.645 2009.647 NA NA  M 24.75546 42.07665 NA 135  Y  Y
6  9 1987.136 1997.452 2008.286 2012.136 NA NA  F 32.40616 72.67785 NA 145  N  N

      smk  gbmi  galb
1 Never smoker Over-wt <NA>
2 Daily smoker Over-wt <NA>
3 Daily smoker Over-wt <NA>
4 Ex smoker Normal Normo
5 Daily smoker Normal <NA>
6 Daily smoker Obese <NA>

```

```
> addmargins( table( tt<-table(hh$id) ) )
```

```

      1      2      3      4      5      6      7      8      9     10     11     12     13     14     15
15215 1248    664    335    244    146    108    86    46    58    26    23    14    13    10
      16     17     18     19     20     21     22     23     24     25     26     27     28     29     30
      11      8      9      6      5      4      3      3      3      2      5      1      1      1      1
      32     33     36     39     44     46     49     53     56     63     83    Sum
      2      2      1      1      1      2      1      1      1      2      1 18314

```

```
> ( wh <- as.numeric(names(tt[tt==7][1]))+0:5 )
```

```
[1] 11 12 13 14 15 16
```

```
> subset( hh, id %in% wh)[,1:7]
```

```

  id  dob  dodm  doe  dox  doh dod
8  11 1946.444 1965.535 2006.871 2012.445 2002.134 NA
9  11 1946.444 1965.535 2006.871 2012.445 2003.396 NA
10 11 1946.444 1965.535 2006.871 2012.445 2004.297 NA
11 11 1946.444 1965.535 2006.871 2012.445 2004.568 NA
12 11 1946.444 1965.535 2006.871 2012.445 2011.292 NA
13 11 1946.444 1965.535 2006.871 2012.445 2013.086 NA
14 11 1946.444 1965.535 2006.871 2012.445 2013.540 NA
15 12 1967.027 2010.301 2011.027 2012.026      NA NA
16 13 1990.120 1994.999 2009.121 2012.119      NA NA
17 16 1992.919 1996.415 2010.909 2011.851 2003.426 NA
18 16 1992.919 1996.415 2010.909 2011.851 2003.856 NA
19 16 1992.919 1996.415 2010.909 2011.851 2009.948 NA

```

We exclude episodes that follow less than one week (1/52 years) after a previous, as these are considered to be double registrations:

```
> dHH <- c( NA, diff(hh$doh) )
> HHx <- ( !is.na(dHH)      &
+         dHH < 1/52      &
+         duplicated( hh$id ) )
> table( HHx, exclude=NULL )
HHx
FALSE  TRUE  <NA>
26515  2119      0
```

We take a quick look at the results and exclude the duplicate registrations:

```
> ( wh <- hh[HHx,"id"][1:5] )
[1] 23 27 27 32 64
> ( wh <- sort(unique( c(wh,wh+1) ) ) )
[1] 23 24 27 28 32 33 64 65
> subset( cbind( hh[,1:7], HHx ), id %in% wh )
   id    dob    dodm    doe    dox    doh  dod  HHx
23 23 1978.720 2002.246 2011.719 2012.721 2002.260  NA FALSE
24 23 1978.720 2002.246 2011.719 2012.721 2002.296  NA FALSE
25 23 1978.720 2002.246 2011.719 2012.721 2008.344  NA FALSE
26 23 1978.720 2002.246 2011.719 2012.721 2008.344  NA  TRUE
27 24 1965.726 2002.085 2006.665 2012.700      NA  NA FALSE
30 27 1987.380 1998.665 2006.999 2011.908 2000.418  NA FALSE
31 27 1987.380 1998.665 2006.999 2011.908 2000.875  NA FALSE
32 27 1987.380 1998.665 2006.999 2011.908 2000.875  NA  TRUE
33 27 1987.380 1998.665 2006.999 2011.908 2008.656  NA FALSE
34 27 1987.380 1998.665 2006.999 2011.908 2011.183  NA FALSE
35 27 1987.380 1998.665 2006.999 2011.908 2012.412  NA FALSE
36 27 1987.380 1998.665 2006.999 2011.908 2012.710  NA FALSE
37 27 1987.380 1998.665 2006.999 2011.908 2012.793  NA FALSE
38 27 1987.380 1998.665 2006.999 2011.908 2013.061  NA FALSE
39 27 1987.380 1998.665 2006.999 2011.908 2013.061  NA  TRUE
40 28 1971.793 1995.999 2009.795 2012.795      NA  NA FALSE
42 32 1985.666 1997.001 2007.158 2012.667 2009.206  NA FALSE
43 32 1985.666 1997.001 2007.158 2012.667 2009.206  NA  TRUE
73 64 1950.778 1996.453 2011.618 2012.453 1997.559  NA FALSE
74 64 1950.778 1996.453 2011.618 2012.453 1997.559  NA  TRUE
75 65 1956.804 1977.001 2007.158 2012.804      NA  NA FALSE

> hh <- hh[!HHx,]
> dim( hh )
[1] 26515    17
```

The date `dox` is the date of the last clinical record, but patients are followed for death only till the end of 2011:

```
> range( hh$doe, na.rm=TRUE )
[1] 2005.335 2012.705
> range( hh$dox, na.rm=TRUE )
[1] 2006.534 2012.998
> range( hh$doh, na.rm=TRUE )
[1] 1985.132 2013.855
> range( hh$dod, na.rm=TRUE )
[1] 2006.977 2011.990
```

We deduce that follow-up for some start at 2005-05-01, but mostly after 2006:

```
> cal.yr( as.Date( "2005-05-01" ) )
[1] 2005.329
attr(,"class")
[1] "cal.yr" "numeric"
> par( mfrow=c(1,3) )
> with( hh, hist(doe, col=1, breaks=seq(2005,2013,1/12),
+          xaxt="n", xlab="Date of entry", ylim=c(0,2000) ) )
> abline(v=2005:2013,col=2)
> axis( side=1, at=2005:2013, labels=NA )
> axis( side=1, at=2005+0.5+0:3*2, labels=2005+0:3*2, tcl=0 )
> with( hh, hist(dod, col=1, breaks=seq(2005,2013,1/12),
+          xaxt="n", xlab="Date of death", ylim=c(0,2000) ) )
> abline(v=2005:2013,col=2)
> axis( side=1, at=2005:2013, labels=NA )
> axis( side=1, at=2005+0.5+0:3*2, labels=2005+0:3*2, tcl=0 )
> with( hh, hist(dox, col=1, breaks=seq(2005,2013,1/12),
+          xaxt="n", xlab="Date of exit", ylim=c(0,2000) ) )
> abline(v=2005:2013,col=2)
> axis( side=1, at=2005:2013, labels=NA )
> axis( side=1, at=2005+0.5+0:3*2, labels=2005+0:3*2, tcl=0 )
```

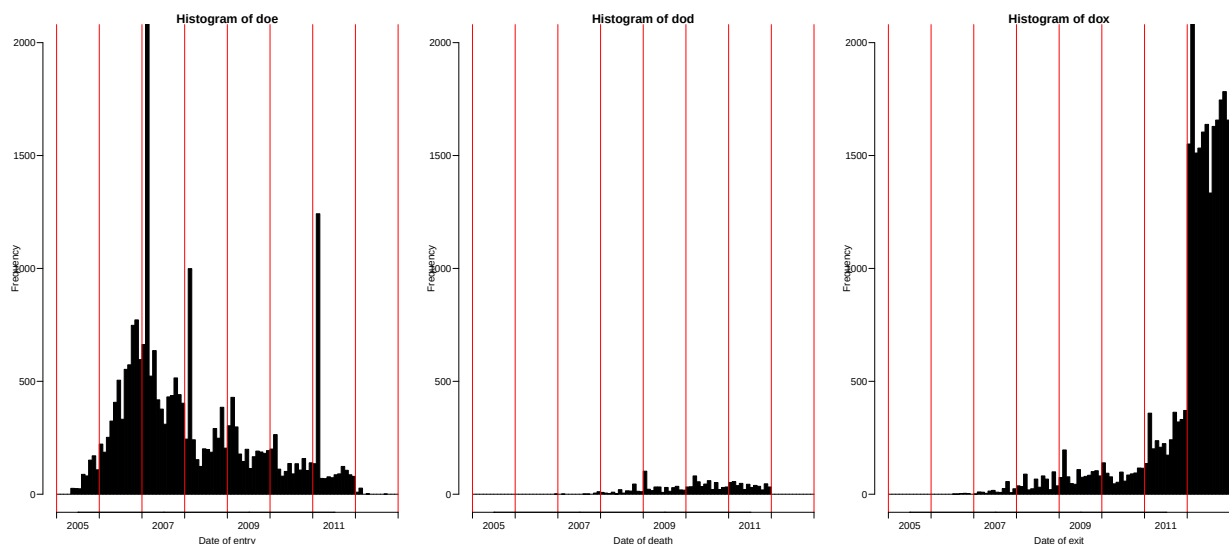


Figure 2: *Date of entry for the persons in the study.*

Thus, if we include persons beyond the date of follow-up for death because of known HH or clinical visit, we get a survival bias, because persons dead after 2011-12-31 without clinical information are not included till their date of death as they should be. Therefore we must censor follow-up at the 2012-01-01 — all persons not dead at this date are under observation for HH only to this date:

```
> hh$doh <- pmin( hh$dod, 2012, na.rm=TRUE )
```

In order to set up a proper `Lexis` object for follow through the states, we must construct a HH counter indicating the state (no. of previous HHs) the person is in at `doh` — essentially the number of non-missing HH dates at `doh`. This variable will be 0 for those with missing `doh`. It is used to exclude records referring to episodes beyond 6, and we only also include dates of HH before `doh` (but always the first record for each person):

```

> hh$nhh <- ave( hh$doh, hh$id, FUN=function(x) cumsum(!is.na(x)) )
> length( unique( hh$id ) )
[1] 18314

> hh <- subset( hh, nhh < 7 )
> length( unique( hh$id ) )
[1] 18314

> subset( hh, id %in% wh)[,-(8:15)]
   id   dob   dodm   doe  dox   doh  dod   gbmi  galb  nhh
23 23 1978.720 2002.246 2011.719 2012 2002.260 NA   Normal  <NA>   1
24 23 1978.720 2002.246 2011.719 2012 2002.296 NA   Normal  <NA>   2
25 23 1978.720 2002.246 2011.719 2012 2008.344 NA   Normal  <NA>   3
27 24 1965.726 2002.085 2006.665 2012      NA   NA   Normal Normo   0
30 27 1987.380 1998.665 2006.999 2012 2000.418 NA Under-wt Normo   1
31 27 1987.380 1998.665 2006.999 2012 2000.875 NA Under-wt Normo   2
33 27 1987.380 1998.665 2006.999 2012 2008.656 NA Under-wt Normo   3
34 27 1987.380 1998.665 2006.999 2012 2011.183 NA Under-wt Normo   4
35 27 1987.380 1998.665 2006.999 2012 2012.412 NA Under-wt Normo   5
36 27 1987.380 1998.665 2006.999 2012 2012.710 NA Under-wt Normo   6
40 28 1971.793 1995.999 2009.795 2012      NA   NA   Over-wt Normo   0
42 32 1985.666 1997.001 2007.158 2012 2009.206 NA   Obese Normo   1
73 64 1950.778 1996.453 2011.618 2012 1997.559 NA   Obese Normo   1
75 65 1956.804 1977.001 2007.158 2012      NA   NA   Normal Normo   0

```

The relevant follow-up is then in the records from `doe` to `dox`, subdivided at `doh` with the state accordingly updated. We cannot construct the `Lexis` object directly from the records, because persons with say H dates of HH in the interval from `doe` to `dox` would need $H + 1$ follow-up records.

Thus we first set up the dataset to form the entire follow-up from `doe` to `dox`, thus just use the first record for each person (selected by `!duplicated`), ignoring any HH information. For baseline calculations etc. we keep a copy `L0`, while `Lh` is the one to modify by intermediate states of HH:

```

> L0 <-
+ Lh <- Lexis( entry = list( per = doe,
+                             age = doe-dob,
+                             dur = doe-dodm ),
+             exit = list( per = dox),
+             entry.status = "No HH",
+             exit.status = ifelse( !is.na(dod) & abs(dod-dox)<0.01,
+                                   "Dead", "No HH" ),
+             id = id,
+             data = subset( hh, !duplicated(id) ),
+             keep = TRUE )
Incompatible factor levels in entry.status and exit.status:
both lex.Cst and lex.Xst now have levels:
No HH Dead

```

We asked `Lexis` to keep the dropped records, so we can list them:

```

> attr( "Lh", "dropped" )[,-(8:15)]
   id   dob   dodm   doe  dox   doh  dod   gbmi  galb  nhh
567  460 1966.898 1976.453 2012.286 2012      NA   NA   Normal Normo   0
1468 1205 1974.189 1994.000 2012.070 2012      NA   NA   Over-wt Normo   0
2503 2026 1992.549 2005.001 2012.100 2012 2011.632 NA   Normal Normo   1
2528 2044 1986.353 2011.999 2012.127 2012      NA   NA   Over-wt  <NA>   0
3568 2918 1972.431 2011.580 2012.103 2012      NA   NA   Normal Normo   0
3572 2925 1970.307 2011.451 2012.122 2012      NA   NA   Normal Normo   0
4194 3402 1993.833 2011.084 2012.040 2012      NA   NA   Normal  <NA>   0

```

```

5987 4727 1979.117 2011.495 2012.157 2012 NA NA Normal Normo 0
7350 5804 1981.332 2011.580 2012.023 2012 NA NA Over-wt <NA> 0
8435 6743 1971.292 1983.999 2012.119 2012 NA NA Over-wt Normo 0
8836 7075 1937.677 1992.453 2012.122 2012 2009.403 NA Under-wt Normo 1
9670 7775 1968.355 1980.453 2012.286 2012 NA NA Over-wt Normo 0
10486 8490 1926.323 1980.453 2012.122 2012 NA NA Normal Micro 0
10719 8696 1992.965 2003.999 2012.160 2012 NA NA Normal Normo 0
12293 9980 1993.890 2010.999 2012.084 2012 NA NA Normal Normo 0
13830 11222 1954.862 1997.452 2012.122 2012 NA NA Over-wt Normo 0
14092 11429 1964.571 2011.451 2012.122 2012 NA NA Over-wt Normo 0
14434 11685 1980.146 1990.452 2012.122 2012 2002.548 NA Normal <NA> 1
15063 12197 1993.537 1998.665 2012.160 2012 NA NA Over-wt Normo 0
16957 13659 1993.439 2005.581 2012.064 2012 NA NA Over-wt Normo 0
17164 13849 1969.231 1997.452 2012.037 2012 NA NA Over-wt Normo 0
17510 14168 1965.507 1981.001 2012.048 2012 2002.939 NA Over-wt <NA> 1
18676 14973 1978.392 2009.162 2012.127 2012 NA NA Over-wt Normo 0
19565 15732 1990.619 2011.451 2012.122 2012 NA NA Under-wt Normo 0
20446 16492 1940.798 1999.999 2012.045 2012 NA NA Normal Normo 0
25783 20934 1926.000 1992.453 2012.122 2012 NA NA Normal Micro 0
26166 21211 1989.874 2002.452 2012.705 2012 NA NA Obese Normo 0
27367 22178 1942.553 1998.413 2012.048 2012 2003.856 NA Over-wt Normo 1
28140 22861 1943.372 1947.999 2012.160 2012 2011.090 NA Under-wt Normo 1

```

...so we see that they are all persons entering after the revised `doe` which was constrained to be maximally equal to 2012.0.

We list original records and corresponding `Lexis` records for 8 persons:

```

> subset( hh, id %in% wh ),-(8:15)]
  id  dob  dodm  doe  dox  doh  dod  gbmi  galb  nhh
23 23 1978.720 2002.246 2011.719 2012 2002.260 NA Normal <NA> 1
24 23 1978.720 2002.246 2011.719 2012 2002.296 NA Normal <NA> 2
25 23 1978.720 2002.246 2011.719 2012 2008.344 NA Normal <NA> 3
27 24 1965.726 2002.085 2006.665 2012 NA NA Normal Normo 0
30 27 1987.380 1998.665 2006.999 2012 2000.418 NA Under-wt Normo 1
31 27 1987.380 1998.665 2006.999 2012 2000.875 NA Under-wt Normo 2
33 27 1987.380 1998.665 2006.999 2012 2008.656 NA Under-wt Normo 3
34 27 1987.380 1998.665 2006.999 2012 2011.183 NA Under-wt Normo 4
35 27 1987.380 1998.665 2006.999 2012 2012.412 NA Under-wt Normo 5
36 27 1987.380 1998.665 2006.999 2012 2012.710 NA Under-wt Normo 6
40 28 1971.793 1995.999 2009.795 2012 NA NA Over-wt Normo 0
42 32 1985.666 1997.001 2007.158 2012 2009.206 NA Obese Normo 1
73 64 1950.778 1996.453 2011.618 2012 1997.559 NA Obese Normo 1
75 65 1956.804 1977.001 2007.158 2012 NA NA Normal Normo 0

> subset( Lh, lex.id %in% wh ),[1:14]
  per  age  dur  lex.dur  lex.Cst  lex.Xst  lex.id  id  dob  dodm
23 2011.719 32.99932 9.472964 0.2806297 No HH No HH 23 23 1978.720 2002.246
27 2006.665 40.93908 4.580424 5.3347023 No HH No HH 24 24 1965.726 2002.085
30 2006.999 19.61944 8.334018 5.0006845 No HH No HH 27 27 1987.380 1998.665
40 2009.795 38.00137 13.796030 2.2053388 No HH No HH 28 28 1971.793 1995.999
42 2007.158 21.49213 10.157426 4.8418891 No HH No HH 32 32 1985.666 1997.001
73 2011.618 60.84052 15.164956 0.3819302 No HH No HH 64 64 1950.778 1996.453
75 2007.158 50.35455 30.157426 4.8418891 No HH No HH 65 65 1956.804 1977.001

  doe  dox  doh  dod
23 2011.719 2012 2002.260 NA
27 2006.665 2012 NA NA
30 2006.999 2012 2000.418 NA
40 2009.795 2012 NA NA
42 2007.158 2012 2009.206 NA
73 2011.618 2012 1997.559 NA
75 2007.158 2012 NA NA

> summary( Lh )

```

Transitions:

```

To
From      No HH Dead  Records:  Events: Risk time:  Persons:
No HH 17612  673      18285      673   70002.17      18285

```

Once this dataset has been set up, we must cut the follow-up at the dates of HH, `doh`, records for which can be defined in two different ways:

```

> with( hh, table( nhh>0, is.na(doh), useNA="ifany" ) )
      FALSE TRUE
FALSE      0 12440
TRUE    12246      0

```

We set up a data frame with the relevant information to define transitions between states.

```

> hc <- subset( hh, nhh>0 )[,c("id","doh","nhh")]
> names( hc ) <- c("lex.id","cut","new.state")
> hc$new.state <- factor( hc$new.state,
+                          levels = 1:6,
+                          labels = c("One",
+                                     "Two",
+                                     "Three",
+                                     "Four",
+                                     "Five",
+                                     ">Six") )

```

Look at the records in the `Lexis` object and the records for the same persons in the `cut` data frame:

```

> wh <- unique( hc$lex.id )[1:4]
> subset( Lh, lex.id %in% wh )[,1:14]

```

	per	age	dur	lex.dur	lex.Cst	lex.Xst	lex.id	id	dob	dodm
8	2006.871	60.42710	41.336071	5.1293634	No HH	No HH	11 11	1946.444	1965.535	
17	2010.909	17.99042	14.494182	1.0910335	No HH	No HH	16 16	1992.919	1996.415	
20	2011.158	43.40589	21.158111	0.8418891	No HH	No HH	17 17	1967.752	1990.000	
21	2007.999	56.88980	7.915127	4.0013689	No HH	No HH	19 19	1951.109	2000.084	

```

      doe dox      doh dod
8  2006.871 2012 2002.134  NA
17 2010.909 2012 2003.426  NA
20 2011.158 2012 1999.281  NA
21 2007.999 2012 2013.639  NA
> subset( hc, lex.id %in% wh )[,-(8:15)]

```

	lex.id	cut	new.state
8	11	2002.134	One
9	11	2003.396	Two
10	11	2004.297	Three
11	11	2004.568	Four
12	11	2011.292	Five
13	11	2013.086	>Six
17	16	2003.426	One
18	16	2003.856	Two
19	16	2009.948	Three
20	17	1999.281	One
21	19	2013.639	One

The following should necessarily be a table of 0s and 1s, and the second the distribution of HH dates by type:

```

> table( tt <- with( hc, table( lex.id, new.state ) ) )

```

```

      0      1
22998 12246
> apply( tt, 2, sum )

      One   Two Three   Four   Five   >Six
5874   2751  1558    945    655    463

```

Not that this is the distribution of HH *dates* not of HH *events*, dates before date of entry (doh<doe) do not represent events; they are only used to update the state at entry.

Thus we have transitions to One, Two, ... HHs — note that `cutLexis` will update the current state even if the transition is *prior* to the entry date. Since `cutLexis` (currently) only accept one record per person we will have to do the updating in a loop over the levels present in the `new.state` variable:

```

> lv <- levels( hc$new.state )
> nl <- nlevels( hc$new.state )
> for( nx in 1:nl )
+ {
+   cat( "New state", nx, lv[nx], "\n" )
+   Lh <- cutLexis( Lh, cut = subset( hc, new.state == lv[nx] ),
+                 precursor.states = levels(Lh)[1:nx],
+                 new.scale = paste("hh",nx,sep="") )
+   cat( "levels(Lh)=", paste(levels(Lh),collapse="," ) )
+   cat( "; nrow(Lh)=", nrow(Lh), "\n" )
+ }

New state 1 One
levels(Lh)= No HH, One, Dead; nrow(Lh)= 19167
New state 2 Two
levels(Lh)= No HH, One, Two, Dead; nrow(Lh)= 19716
New state 3 Three
levels(Lh)= No HH, One, Two, Three, Dead; nrow(Lh)= 20083
New state 4 Four
levels(Lh)= No HH, One, Two, Three, Four, Dead; nrow(Lh)= 20334
New state 5 Five
levels(Lh)= No HH, One, Two, Three, Four, Five, Dead; nrow(Lh)= 20516
New state 6 >Six
levels(Lh)= No HH, One, Two, Three, Four, Five, >Six, Dead; nrow(Lh)= 20654

> summary( Lh, t=T )

Transitions:
  To
From  No HH  One  Two Three Four Five >Six Dead  Records:  Events: Risk time:  Persons:
No HH 12425  882    0    0    0    0    0  320    13627    1202   49663.18    13627
One    0 2860  549    0    0    0    0  143    3552    692   11449.02    3552
Two    0    0 1063  367    0    0    0    60    1490    427   4089.07    1490
Three  0    0    0  514  251    0    0    54    819    305   1972.07    819
Four   0    0    0    0  243  182    0    34    459    216    976.01    459
Five   0    0    0    0    0  154  138   11    303    149    542.81    303
>Six   0    0    0    0    0    0  353   51    404     51   1310.00    404
Sum   12425 3742 1612  881  494  336  491  673   20654   3042  70002.17   18285

Time scales:  time.scale time.since
1           per
2           age
3           dur
4          hh1          One
5          hh2          Two
6          hh3          Three
7          hh4          Four
8          hh5          Five
9          hh6          >Six

> summary( Lh, by=Lh$sex )

```

\$M

Transitions:

From	To	No HH	One	Two	Three	Four	Five	>Six	Dead	Records:	Events:	Risk time:	Persons:
No HH	No HH	6911	515	0	0	0	0	0	202	7628	717	27344.21	7628
	One	0	1672	329	0	0	0	0	94	2095	423	6723.63	2095
	Two	0	0	655	225	0	0	0	32	912	257	2502.90	912
	Three	0	0	0	324	161	0	0	35	520	196	1247.71	520
	Four	0	0	0	0	150	124	0	20	294	144	615.36	294
	Five	0	0	0	0	0	99	84	8	191	92	337.99	191
	>Six	0	0	0	0	0	0	229	33	262	33	853.57	262
Sum		6911	2187	984	549	311	223	313	424	11902	1862	39625.35	10464

\$F

Transitions:

From	To	No HH	One	Two	Three	Four	Five	>Six	Dead	Records:	Events:	Risk time:	Persons:
No HH	No HH	5514	367	0	0	0	0	0	118	5999	485	22318.98	5999
	One	0	1188	220	0	0	0	0	49	1457	269	4725.39	1457
	Two	0	0	408	142	0	0	0	28	578	170	1586.18	578
	Three	0	0	0	190	90	0	0	19	299	109	724.36	299
	Four	0	0	0	0	93	58	0	14	165	72	360.65	165
	Five	0	0	0	0	0	55	54	3	112	57	204.82	112
	>Six	0	0	0	0	0	0	124	18	142	18	456.43	142
Sum		5514	1555	628	332	183	113	178	249	8752	1180	30376.82	7821

We can then illustrate how the transitions between different levels of HH play out over the period:

```
> al <- c(20,17.9,15.1)
> ang <- pi*c(al,11,22-rev(al))/22
> bp <- list( x=c( cos(ang)* 45+50,50),
+             y=c( sin(ang)*105-10,10) )
> msbx <-
+ boxes( Lh, boxpos=bp, hm=1.2, wm=1.1,
+         show.BE="nz", scale.R=100, lwd=4, lwd.arr=4,
+         col.txt=gray(c(0,1))[c(rep(1,7),2)],
+         col.bg =gray(c(1,0.5))[c(rep(1,7),2)],
+         col.arr=gray(c(0,0.5))[c(1,2,1,2,1,2,1,2,1,2,2)] )
```

Also note that we have defined 6 new time scales in order to address the effect of time since previous HH event:

```
> cbind( attr(Lh,"time.scales"),
+        attr(Lh,"time.since") )
      [,1] [,2]
[1,] "per" ""
[2,] "age" ""
[3,] "dur" ""
[4,] "hh1" "One"
[5,] "hh2" "Two"
[6,] "hh3" "Three"
[7,] "hh4" "Four"
[8,] "hh5" "Five"
[9,] "hh6" ">Six"
```

A look at how these play out in the real dataset, compared to the input records from the cut dataset:

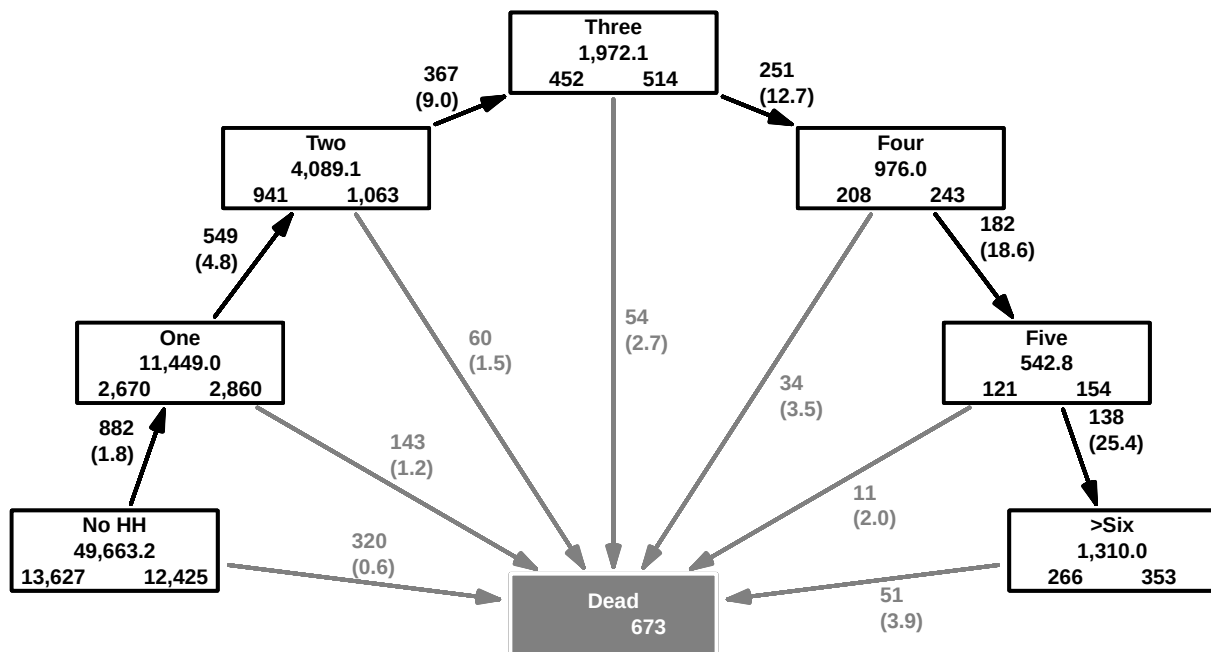


Figure 3: Transitions between states — number of recorded HHs since 1995. The follow-up is from 2005-03-01 through 2011-12-31. The number in the middle of each box is the number of person-years, the two numbers at the bottom of each box are the number of persons starting, resp. ending their follow-up in each state. The numbers on the arrows are the number of transitions and transition rates in events per 100 years (% per year).

```
> wh <- unique( Lh[Lh$lex.Cst==">Six","lex.id"] )
> cl <- c( timeScales( Lh ),
+         names( Lh )[grep("lex",names(Lh))],
+         "doe","dox","doh" )
> subset( Lh, lex.id %in% wh[1:4] )[,cl]
```

	per	age	dur	hh1	hh2	hh3	hh4	hh5
128	2006.832	51.92334	50.83368	9.303217	1.322382	NA	NA	NA
129	2008.251	53.34155	52.25188	10.721424	2.740589	0.0000000	NA	NA
130	2008.275	53.36619	52.27652	10.746064	2.765229	0.02464066	0.0000000	NA
131	2009.403	54.49418	53.40452	11.874059	3.893224	1.15263518	1.1279945	0.0000000
20647	2011.396	56.48734	55.39767	13.867214	5.886379	3.14579055	3.1211499	1.9931554
181	2006.496	27.59480	11.66324	8.027379	6.918549	3.43874059	0.7364819	0.2765229
20697	2007.372	28.47091	12.53936	8.903491	7.794661	4.31485284	1.6125941	1.1526352
20729	2006.747	50.55989	26.74880	5.103354	3.356605	3.30732375	2.2012320	0.7967146
20741	2011.561	40.99932	11.56194	11.474333	9.448323	9.05954825	9.0075291	8.6680356

```

      hh6    lex.dur lex.Cst lex.Xst lex.id    doe    dox    doh
128      NA  1.41820671    Two   Three   139 2006.832 2012 1997.529
129      NA  0.02464066   Three   Four   139 2006.832 2012 1997.529
130      NA  1.12799452   Four    Five   139 2006.832 2012 1997.529
131      NA  1.99315537   Five   >Six   139 2006.832 2012 1997.529
20647 0.0000000 0.60369610 >Six   >Six   139 2006.832 2012 1997.529
181      NA  0.87611225   Five   >Six   201 2006.496 2012 1998.468
20697 0.0000000 4.62833676 >Six   >Six   201 2006.496 2012 1998.468
20729 0.7583847 5.25256674 >Six   >Six   237 2006.747 2012 2001.644
20741 8.4818617 0.43942505 >Six   >Six   250 2011.561 2012 2000.086

> subset( hc, lex.id %in% wh[1:4] )
      lex.id    cut new.state
162     139 1997.529      One
```

```

163      139 2005.510      Two
164      139 2008.251      Three
165      139 2008.275      Four
166      139 2009.403      Five
167      139 2011.396      >Six
234      201 1998.468      One
235      201 1999.577      Two
236      201 2003.057      Three
237      201 2005.759      Four
238      201 2006.219      Five
239      201 2007.372      >Six
276      237 2001.644      One
277      237 2003.391      Two
278      237 2003.440      Three
279      237 2004.546      Four
280      237 2005.951      Five
281      237 2005.989      >Six
294      250 2000.086      One
295      250 2002.112      Two
297      250 2002.501      Three
299      250 2002.553      Four
301      250 2002.893      Five
303      250 2003.079      >Six

```

Analysis of HH rates will only include persons that have a recorded systolic blood pressure, and have at most 5 recorded HHs — that is the last possible transition is from “Five” to “>Six”, so any follow-up in the “>Six” state should be disregarded, since transitions *out* of this state are not modeled. Hence we restrict the dataset to these persons:

```
> summary( subset( Lh, !is.na( sbp ) ) )
```

Transitions:

From	To	No HH	One	Two	Three	Four	Five	>Six	Dead	Records:	Events:	Risk time:	Persons:
No HH	No HH	11844	851	0	0	0	0	0	315	13010	1166	47223.92	13010
	One	0	2754	521	0	0	0	0	138	3413	659	10998.20	3413
	Two	0	0	1016	345	0	0	0	60	1421	405	3891.96	1421
	Three	0	0	0	490	240	0	0	51	781	291	1890.37	781
	Four	0	0	0	0	234	176	0	34	444	210	944.14	444
	Five	0	0	0	0	0	150	133	11	294	144	527.63	294
	>Six	0	0	0	0	0	0	345	50	395	50	1281.94	395
Sum		11844	3605	1537	835	474	326	478	659	19758	2925	66758.16	17492

```

> al <- c(20,17.9,15.1)
> ang <- pi*c(al,11,22-rev(al))/22
> bp <- list( x=c( cos(ang)* 45+50,50),
+             y=c( sin(ang)*105-10,10) )
> msbr <-
+ boxes( subset( Lh, !is.na( sbp ) ),
+        boxpos=bp, hm=1.2, wm=1.1,
+        show.BE="nz", scale.R=100, lwd=4, lwd.arr=4,
+        col.txt=gray(c(0,1))[c(rep(1,7),2)],
+        col.bg =gray(c(1,0.5))[c(rep(1,7),2)],
+        col.arr=gray(c(0,0.5))[c(1,2,1,2,1,2,1,2,1,2,2)] )

```

Finally, we save the datasets for analysis — note we carried the missing `sbp` values over in `Lh`:

```
> save( L0, Lh, file = "../data/Lh.Rda" )
```

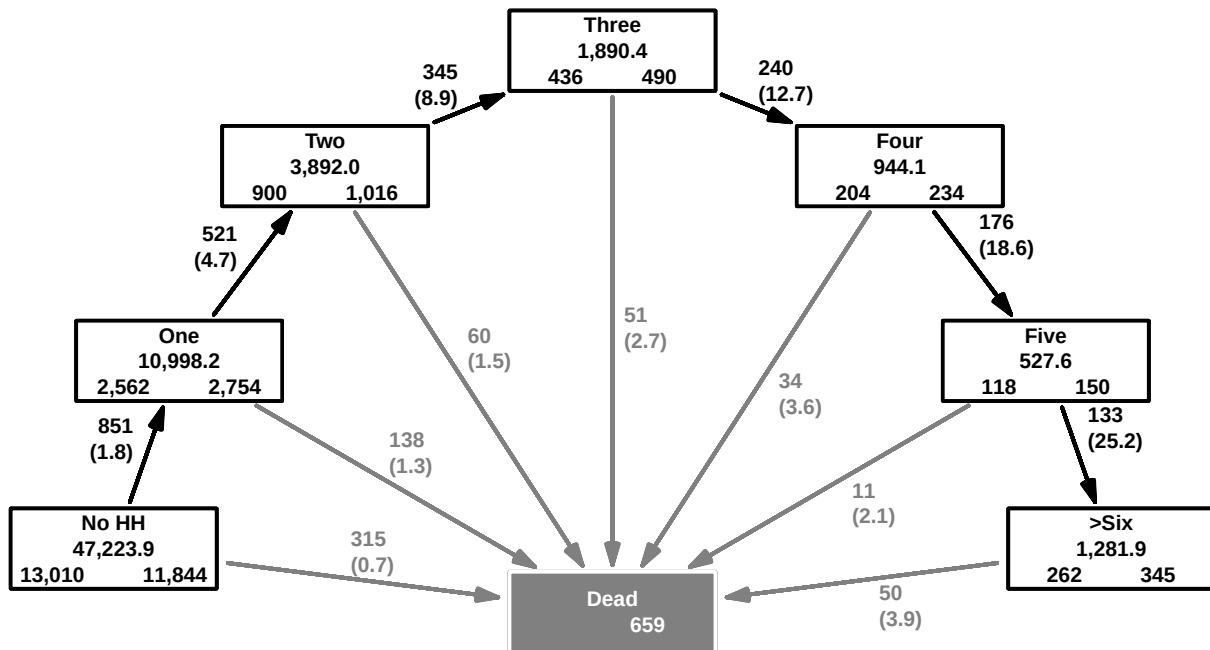


Figure 4: *Transitions between states — number of recorded HHs since 1995 in the analysis dataset where all persons have all clinical measurements (except possibly U-albumin). The follow-up is from 2005-03-01 through 2011-12-31. The number in the middle of each box is the number of person-years, the two numbers at the bottom of each box are the number of persons starting, resp. ending their follow-up in each state. The numbers on the arrows are the number of transitions and transition rates in events per 100 years (% per year).*

4 Analysis of HH rates

In the introduction we set up the following model for the HH rates:

$$\log(\lambda) = X\beta + f(a) + g(d) + k(t) + \ell_H(h)$$

with $X\beta$ representing effects of the baseline clinical variables on the outcome, and the remaining terms representing the different time scales, of which calendar time is of primary interest.

In order to conduct the analysis we split the dataset in 3-month intervals along a suitable time-scale — we use calendar time, `per`, the first (which default):

```
> library( Epi )
> library( splines )
> sessionInfo()
R version 3.3.1 (2016-06-21)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 14.04.5 LTS

locale:
 [1] LC_CTYPE=en_US.UTF-8      LC_NUMERIC=C               LC_TIME=en_US.UTF-8
 [4] LC_COLLATE=en_US.UTF-8    LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8      LC_NAME=C                  LC_ADDRESS=C
[10] LC_TELEPHONE=C            LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
```

attached base packages:

```
[1] splines      utils      datasets  graphics  grDevices  stats      methods    base
```

other attached packages:

```
[1] Epi_2.6
```

loaded via a namespace (and not attached):

```
[1] cmprsk_2.2-7      MASS_7.3-45      Matrix_1.2-6     plyr_1.8.4
[5] parallel_3.3.1    survival_2.39-5  etm_0.6-2        Rcpp_0.12.5
[9] grid_3.3.1        numDeriv_2014.2-1 lattice_0.20-33
```

```
> load( file = "../data/Lh.Rda" )
```

```
> summary( L0 )
```

Transitions:

To

From	No HH	No HH	Dead	Records:	Events:	Risk time:	Persons:
No HH	17612	673		18285	673	70002.17	18285

```
> summary( Lh, timeScales=T )
```

Transitions:

To

From	No HH	One	Two	Three	Four	Five	>Six	Dead	Records:	Events:	Risk time:	Persons:
No HH	12425	882	0	0	0	0	0	320	13627	1202	49663.18	13627
One	0	2860	549	0	0	0	0	143	3552	692	11449.02	3552
Two	0	0	1063	367	0	0	0	60	1490	427	4089.07	1490
Three	0	0	0	514	251	0	0	54	819	305	1972.07	819
Four	0	0	0	0	243	182	0	34	459	216	976.01	459
Five	0	0	0	0	0	154	138	11	303	149	542.81	303
>Six	0	0	0	0	0	0	353	51	404	51	1310.00	404
Sum	12425	3742	1612	881	494	336	491	673	20654	3042	70002.17	18285

Time scales: time.scale time.since

1	per	
2	age	
3	dur	
4	hh1	One
5	hh2	Two
6	hh3	Three
7	hh4	Four
8	hh5	Five
9	hh6	>Six

```
> system.time(
```

```
+ Sh <- splitLexis( Lh, breaks=seq(1990,2020,1/4) ) )
```

	user	system	elapsed
	2.969	0.064	3.033

```
> summary( Sh, t=T )
```

Transitions:

To

From	No HH	One	Two	Three	Four	Five	>Six	Dead	Records:	Events:	Risk time:
No HH	205326	882	0	0	0	0	0	320	206528	1202	49663.18
One	0	47320	549	0	0	0	0	143	48012	692	11449.02
Two	0	0	16925	367	0	0	0	60	17352	427	4089.07
Three	0	0	0	8165	251	0	0	54	8470	305	1972.07
Four	0	0	0	0	4042	182	0	34	4258	216	976.01
Five	0	0	0	0	0	2255	138	11	2404	149	542.81
>Six	0	0	0	0	0	0	5422	51	5473	51	1310.00
Sum	205326	48202	17474	8532	4293	2437	5560	673	292497	3042	70002.17

Transitions:

To

```

From           Persons:
  No HH         13627
  One           3552
  Two           1490
  Three          819
  Four           459
  Five           303
  >Six           404
  Sum           18285

Time scales:  time.scale time.since
1             per
2             age
3             dur
4             hh1             One
5             hh2             Two
6             hh3             Three
7             hh4             Four
8             hh5             Five
9             hh6             >Six

```

5 Overall rates by calendar time

We can model the overall occurrence rates of HH by lumping together any occurrence of HH in the `Lexis` object, and only counting *new transitions* — we also make a check that we got it right:

```
> ( HHx <- levels( Sh )[2:7] )
[1] "One"    "Two"    "Three"  "Four"   "Five"   ">Six"

> Sh <- transform( Sh, anyHH = lex.Xst %in% HHx & lex.Xst != lex.Cst )
> with( Sh, ftable( lex.Xst, lex.Cst, anyHH, row.vars=3:2 ) )
```

		lex.Xst	No HH	One	Two	Three	Four	Five	>Six	Dead
anyHH	lex.Cst									
FALSE	No HH		205326	0	0	0	0	0	0	320
	One		0	47320	0	0	0	0	0	143
	Two		0	0	16925	0	0	0	0	60
	Three		0	0	0	8165	0	0	0	54
	Four		0	0	0	0	4042	0	0	34
	Five		0	0	0	0	0	2255	0	11
	>Six		0	0	0	0	0	0	5422	51
	Dead		0	0	0	0	0	0	0	0
TRUE	No HH		0	882	0	0	0	0	0	0
	One		0	0	549	0	0	0	0	0
	Two		0	0	0	367	0	0	0	0
	Three		0	0	0	0	251	0	0	0
	Four		0	0	0	0	0	182	0	0
	Five		0	0	0	0	0	0	138	0
	>Six		0	0	0	0	0	0	0	0
	Dead		0	0	0	0	0	0	0	0

As a very crude model we compute the overall HH rates by age and calendar year:

[illegible]

```

      12.5%    37.5%    62.5%    87.5%
2007.640 2009.003 2010.214 2011.415

> map <- glm( anyHH ~ -1 + Ns( age, knots=a.kn, intercept=T ) +
+             Ns( per, knots=p.kn, ref=2010 ),
+             offset = log(lex.dur/100),
+             family = poisson,
+             data = Sh )
> a.pr <- seq(20,90,,200)
> p.pr <- seq(2005,2012,,200)
> rate.a <- ci.exp( map, subset="age",
+                  ctr.mat = Ns( a.pr, knots=a.kn, intercept=T ) )
> rate.p <- ci.exp( map, subset="per",
+                  ctr.mat = Ns( p.pr, knots=p.kn, ref=2010 ) )

```

We now have the predicted rates of “any HH” by age for 2010, and the RR relative to 2010:

```

> par( mfrow=c(1,2) )
> matplot( a.pr, rate.a,
+          type="l", lty=1, lwd=c(3,1,1), col="black",
+          log="y", xlab="Age", ylim=c(2,20),
+          ylab="Overall incidence rates of HH (cases/100 PY)" )
> axis( side=2, at=c(3:10,15), labels=NA )
> matplot( p.pr, rate.p,
+          type="l", lty=1, lwd=c(3,1,1), col="black",
+          log="y", xlab="Date of follow-up", ylim=c(2,20)/5,
+          ylab="HR of HH relative to 2010" )
> axis( side=2, at=c(c(1:10,15)/10,2:4), labels=NA )
> abline( h=1 )

```

It is of course also possible to derive heavily confounded uninterpretable crude rates by calendar time:

```

> mP <- update(map, . ~ -1 + Ns( per, knots=p.kn, intercept=TRUE ) )
> nd <- data.frame( age = 50,
+                  per = p.pr,
+                  lex.dur = 100 )
> rate.P <- ci.pred( mP, nd )

```

These are most reasonably shown together with period-specific rates using an approximate median age of say 50 years:

```

> rbind(
+ summary( Sh$age ),
+ with( Sh, summary( (age+lex.dur/2)*lex.dur/mean(lex.dur) ) ) )
      Min. 1st Qu. Median Mean 3rd Qu. Max.
[1,] 16.06000  36.94 47.96 48.19  59.46 98.6
[2,]  0.02582  36.20 48.70 48.36  61.21 102.9

> pr50 <- ci.pred( map, nd )
> matplot( p.pr, pr50,
+          type="l", lty=1, lwd=c(3,1,1), col="black",
+          log="y", xlab="Date of follow-up", ylim=c(1.5,7),
+          ylab="Crude rate of HH relative (per 100 PY)" )
> matlines( p.pr, rate.P,
+           type="l", lty=1, lwd=c(3,1,1), col=gray(0.5) )

```

The plot in figure 6 shows that the crude rates are highly misleading; they show a slightly flatter calendar time development in rates than the model, presumably because the age-distribution changes over time. Moreover they are about 1.25 times the predicted rates

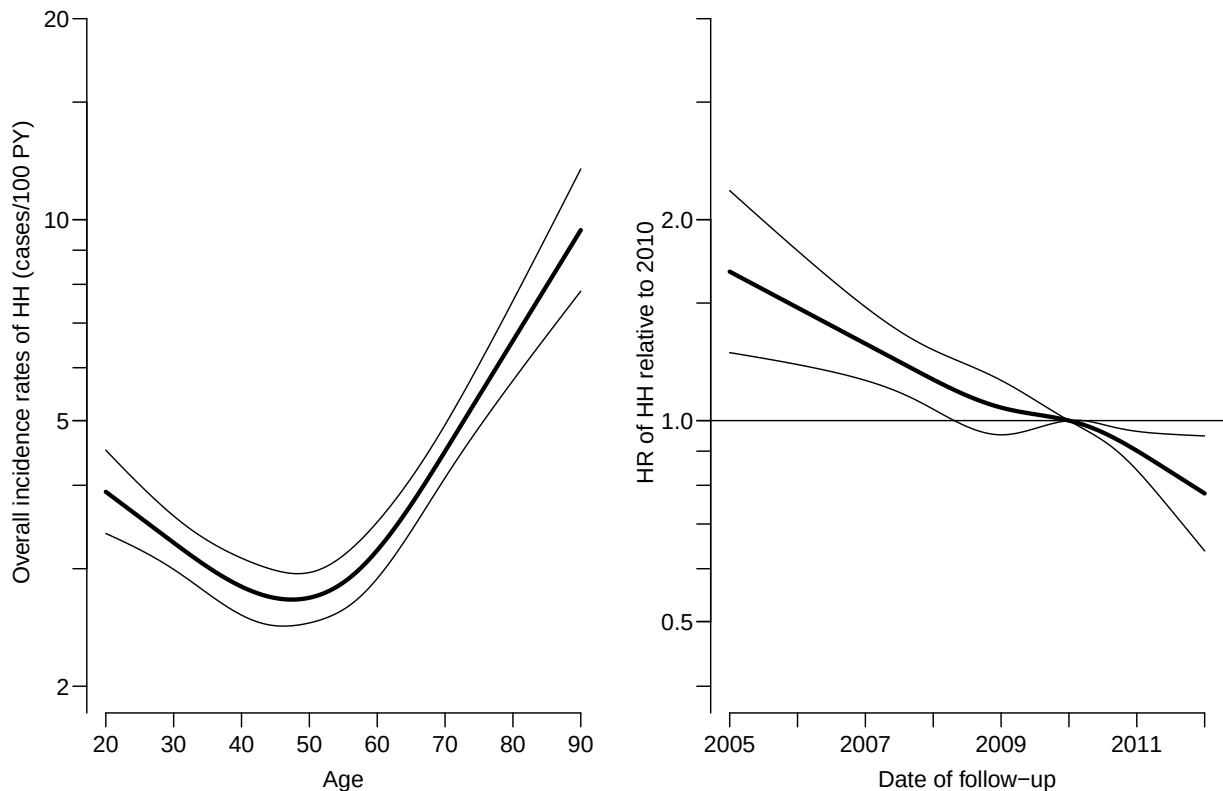


Figure 5: Age-specific incidences in 2010, and period RRs relative to this.

at the median age because of the u-shaped age effect, so the crude rates are more in the ballpark of the rates that would be predicted for persons aged 35 or 65.

We can of course also print select values of these uninterpretable rate-estimates, together with the number of HH events and person-years (1000s):

```
> round( cbind( ppp <- 2005:2011+0.5,
+               xtabs( cbind( HH=anyHH, PY=lex.dur/1000 ) ~ floor(per), data=Sh ),
+               ci.exp( mP, subset="per",
+               ctr.mat = Ns( ppp, knots=p.kn, intercept=TRUE ) ) ), 1 )
```

	HH	PY	exp(Est.)	2.5%	97.5%	
2005	2005.5	5	0.1	5.1	4.0	6.4
2006	2006.5	92	1.8	4.5	3.9	5.2
2007	2007.5	335	8.6	4.0	3.7	4.3
2008	2008.5	453	12.4	3.6	3.3	3.9
2009	2009.5	478	14.4	3.4	3.2	3.6
2010	2010.5	512	15.6	3.2	3.0	3.5
2011	2011.5	494	17.1	2.8	2.6	3.1

```
> # All events and PY by sex:
> round( addmargins( xtabs( cbind( HH=anyHH, PY=lex.dur ) ~ sex,
+                               data=Sh ), margin=1 ), 1 )
```

sex	HH	PY
M	1438.0	39625.4
F	931.0	30376.8
Sum	2369.0	70002.2

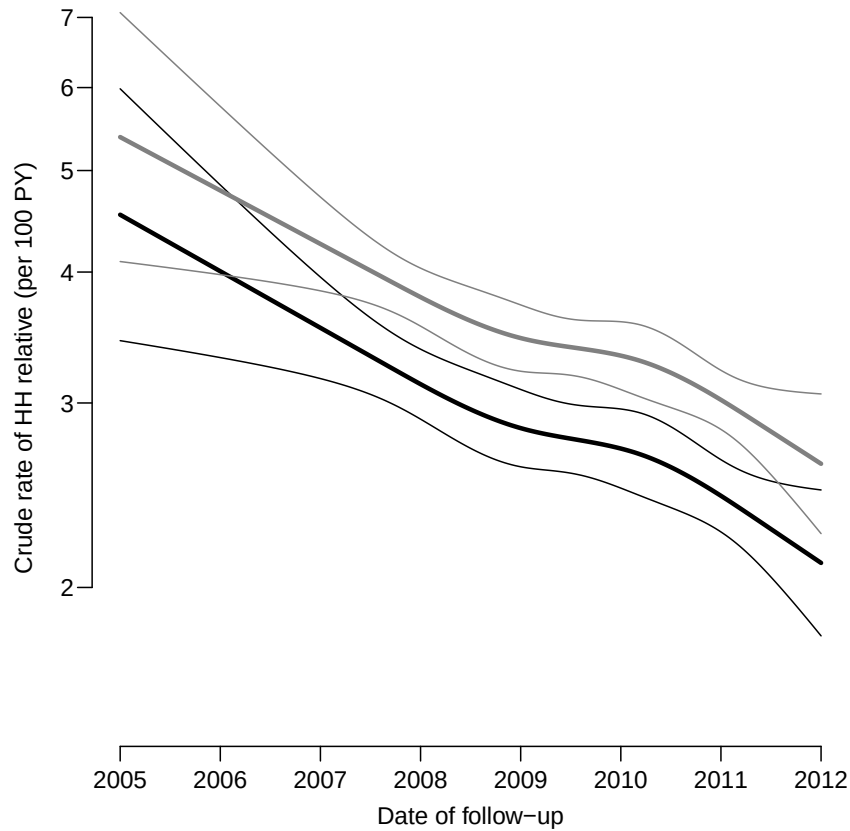


Figure 6: *HH incidence rates for 50-year old patients (black) as predicted from an age-period model, and overall incidence rates as predicted from a confounded period only model (gray).*

6 Analysis by HH status

The modeling of the effect of time since last HH requires a single variable defined for all units in the analysis dataset, this should be equal to the smaller of the variables `hh1` – `hh5`, and 0 for follow-up time before first event. Moreover, we need a definition of the event, which is a new HH, defined by a transition to one of the HH states. Note that this definition of the timescale requires both `pmin` (which leaves those without HH with a NA), and `pmax` (which codes those without HH as 0):

```
> Sh <- transform( Sh, tfh = pmax( 0,
+                               pmin( hh1, hh2, hh3, hh4, hh5, hh6, na.rm=TRUE ),
+                               na.rm=TRUE ),
+                 hh = lex.Xst %in% levels(Sh)[2:7] &
+                 lex.Xst != lex.Cst )
> subset( Sh, lex.id==139 )[ ,c(timeScales(Sh)[- (2:3)], "tfh", "lex.Cst", "lex.Xst", "hh")]
```

	per	hh1	hh2	hh3	hh4	hh5	hh6	tfh
1684	2006.832	9.303217	1.322382	NA	NA	NA	NA	1.32238193
1685	2007.000	9.470910	1.490075	NA	NA	NA	NA	1.49007529
1686	2007.250	9.720910	1.740075	NA	NA	NA	NA	1.74007529
1687	2007.500	9.970910	1.990075	NA	NA	NA	NA	1.99007529
1688	2007.750	10.220910	2.240075	NA	NA	NA	NA	2.24007529
1689	2008.000	10.470910	2.490075	NA	NA	NA	NA	2.49007529
1690	2008.250	10.720910	2.740075	NA	NA	NA	NA	2.74007529
1691	2008.251	10.721424	2.740589	0.00000000	NA	NA	NA	0.00000000

```

1692 2008.275 10.746064 2.765229 0.02464066 0.000000 NA NA 0.00000000
1693 2008.500 10.970910 2.990075 0.24948665 0.224846 NA NA 0.22484600
1694 2008.750 11.220910 3.240075 0.49948665 0.474846 NA NA 0.47484600
1695 2009.000 11.470910 3.490075 0.74948665 0.724846 NA NA 0.72484600
1696 2009.250 11.720910 3.740075 0.99948665 0.974846 NA NA 0.97484600
1697 2009.403 11.874059 3.893224 1.15263518 1.127995 0.00000000 NA 0.00000000
1698 2009.500 11.970910 3.990075 1.24948665 1.224846 0.09685147 NA 0.09685147
1699 2009.750 12.220910 4.240075 1.49948665 1.474846 0.34685147 NA 0.34685147
1700 2010.000 12.470910 4.490075 1.74948665 1.724846 0.59685147 NA 0.59685147
1701 2010.250 12.720910 4.740075 1.99948665 1.974846 0.84685147 NA 0.84685147
1702 2010.500 12.970910 4.990075 2.24948665 2.224846 1.09685147 NA 1.09685147
1703 2010.750 13.220910 5.240075 2.49948665 2.474846 1.34685147 NA 1.34685147
1704 2011.000 13.470910 5.490075 2.74948665 2.724846 1.59685147 NA 1.59685147
1705 2011.250 13.720910 5.740075 2.99948665 2.974846 1.84685147 NA 1.84685147
1706 2011.396 13.867214 5.886379 3.14579055 3.121150 1.99315537 0.0000000 0.00000000
1707 2011.500 13.970910 5.990075 3.24948665 3.224846 2.09685147 0.1036961 0.10369610
1708 2011.750 14.220910 6.240075 3.49948665 3.474846 2.34685147 0.3536961 0.35369610
lex.Cst lex.Xst hh
1684 Two Two FALSE
1685 Two Two FALSE
1686 Two Two FALSE
1687 Two Two FALSE
1688 Two Two FALSE
1689 Two Two FALSE
1690 Two Three TRUE
1691 Three Four TRUE
1692 Four Four FALSE
1693 Four Four FALSE
1694 Four Four FALSE
1695 Four Four FALSE
1696 Four Five TRUE
1697 Five Five FALSE
1698 Five Five FALSE
1699 Five Five FALSE
1700 Five Five FALSE
1701 Five Five FALSE
1702 Five Five FALSE
1703 Five Five FALSE
1704 Five Five FALSE
1705 Five >Six TRUE
1706 >Six >Six FALSE
1707 >Six >Six FALSE
1708 >Six >Six FALSE

```

From the overview of the states in figure 3 we see that the the transition rates modeled only ever involves *one* transition *out* of any state, hence we can model the rates jointly based on the data frame `Sh`. The data stacking commonly used in analysis of multistate models is not needed in this case.

6.1 Clinical variables

The summary of the clinical variables in `L0` (which has one record per person in the follow-up), shows that primarily the `alb` — albuminuria measurements — are prominently missing, whereas only `sbp` among the others have a few missing:

```

> summary.data.frame( L0[,-(1:14)] )
sex      bmi      hba      alb      sbp      aht
M:10464  Min.   :14.45  Min.   : 22.40  Min.   :  0.00  Min.   :  0.0  N:11856
F: 7821  1st Qu.:22.53  1st Qu.: 55.19  1st Qu.:  0.50  1st Qu.:120.0  Y: 6429
        Median :24.74  Median : 63.93  Median :  1.00  Median :130.0
        Mean   :25.36  Mean   : 66.22  Mean   :  7.32  Mean   :132.2

```

```

3rd Qu.:27.47 3rd Qu.: 73.77 3rd Qu.: 2.30 3rd Qu.:141.0
Max. :62.87 Max. :249.73 Max. :1072.03 Max. :250.0
NA's :7126 NA's :793
llt smk gbmi galb nhh
N:12396 Never smoker :6949 Under-wt:1134 Normo:10700 Min. :0.0000
Y: 5889 Daily smoker :5374 Normal :8505 Micro: 415 1st Qu.:0.0000
Ex smoker :2361 Over-wt :6372 Macro: 44 Median :0.0000
No information :2964 Obese :2274 NA's : 7126 Mean :0.3209
Occasional smoker: 637 3rd Qu.:1.0000
Max. :1.0000

```

Since we will be using only models where we include effects of the clinical variables, we exclude those with missing values of `sbp`, 793 persons or 4.3%:

```
> summary.data.frame( subset( L0, !is.na(sbp) ) )
```

```

per age dur lex.dur lex.Cst
Min. :2005 Min. :16.06 Min. : 0.000 Min. :0.004107 No HH:17492
1st Qu.:2007 1st Qu.:34.22 1st Qu.: 7.168 1st Qu.:2.769336 Dead : 0
Median :2007 Median :45.69 Median :16.504 Median :4.420260
Mean :2008 Mean :45.99 Mean :18.945 Mean :3.816497
3rd Qu.:2009 3rd Qu.:57.71 3rd Qu.:28.501 3rd Qu.:4.877481
Max. :2012 Max. :94.00 Max. :75.253 Max. :6.665298

lex.Xst lex.id id dob dodm
No HH:16833 Min. : 1 Min. : 1 Min. :1913 Min. :1933
Dead : 659 1st Qu.: 5756 1st Qu.: 5756 1st Qu.:1950 1st Qu.:1979
Median :11582 Median :11582 Median :1962 Median :1991
Mean :11594 Mean :11594 Mean :1962 Mean :1989
3rd Qu.:17392 3rd Qu.:17392 3rd Qu.:1974 3rd Qu.:2001
Max. :23268 Max. :23268 Max. :1995 Max. :2012

doe dox doh dod sex bmi
Min. :2005 Min. :2007 Min. :1985 Min. :2007 M:10047 Min. :14.45
1st Qu.:2007 1st Qu.:2012 1st Qu.:1998 1st Qu.:2010 F: 7445 1st Qu.:22.53
Median :2007 Median :2012 Median :2003 Median :2010 Median :24.76
Mean :2008 Mean :2012 Mean :2003 Mean :2010 Mean :25.38
3rd Qu.:2009 3rd Qu.:2012 3rd Qu.:2007 3rd Qu.:2011 3rd Qu.:27.47
Max. :2012 Max. :2012 Max. :2014 Max. :2012 Max. :62.87
NA's :11845 NA's :16833

hba alb sbp aht llt
Min. : 22.40 Min. : 0.000 Min. : 0.0 N:11097 N:11632
1st Qu.: 55.19 1st Qu.: 0.500 1st Qu.:120.0 Y: 6395 Y: 5860
Median : 63.93 Median : 1.000 Median :130.0
Mean : 66.15 Mean : 7.308 Mean :132.2
3rd Qu.: 73.77 3rd Qu.: 2.300 3rd Qu.:141.0
Max. :249.73 Max. :1072.031 Max. :250.0
NA's :6424

smk gbmi galb nhh
Never smoker :6891 Under-wt:1078 Normo:10613 Min. :0.0000
Daily smoker :5346 Normal :8106 Micro: 412 1st Qu.:0.0000
Ex smoker :2345 Over-wt :6115 Macro: 43 Median :0.0000
No information :2275 Obese :2193 NA's : 6424 Mean :0.3228
Occasional smoker: 635 3rd Qu.:1.0000
Max. :1.0000

```

```
> summary.data.frame( subset( Sh, !is.na(sbp) ) )
```

```

lex.id per age dur hh1
Min. : 1 Min. :2005 Min. :16.06 Min. : 0.00 Min. : 0.00
1st Qu.: 5788 1st Qu.:2008 1st Qu.:37.02 1st Qu.:10.75 1st Qu.: 4.43
Median :11593 Median :2010 Median :48.09 Median :20.05 Median : 8.12
Mean :11616 Mean :2010 Mean :48.32 Mean :22.13 Mean : 8.07
3rd Qu.:17407 3rd Qu.:2011 3rd Qu.:59.61 3rd Qu.:31.80 3rd Qu.:11.69
Max. :23268 Max. :2012 Max. :98.60 Max. :78.75 Max. :26.62
NA's :196440

```

```

      hh2          hh3          hh4          hh5          hh6
Min.   : 0.00      Min.   : 0.00      Min.   : 0.00      Min.   : 0.00      Min.   : 0.00
1st Qu.: 3.47      1st Qu.: 2.81      1st Qu.: 2.55      1st Qu.: 2.35      1st Qu.: 2.36
Median : 6.53      Median : 5.60      Median : 5.35      Median : 5.28      Median : 5.02
Mean   : 6.77      Mean   : 6.03      Mean   : 5.75      Mean   : 5.59      Mean   : 5.37
3rd Qu.: 9.85      3rd Qu.: 8.90      3rd Qu.: 8.56      3rd Qu.: 8.39      3rd Qu.: 7.93
Max.   :17.65      Max.   :17.48      Max.   :17.37      Max.   :16.85      Max.   :16.79
NA's   :242557     NA's   :259076     NA's   :267192     NA's   :271312     NA's   :273646

      lex.dur      lex.Cst      lex.Xst      id      dob
Min.   :0.0001711 No HH :196440 No HH :195274 Min.   : 1 Min.   :1913
1st Qu.:0.2500000 One   : 46117 One   : 46309 1st Qu.: 5788 1st Qu.:1950
Median :0.2500000 Two   : 16519 Two   : 16635 Median :11593 Median :1961
Mean   :0.2392731 Three  :  8116 Three  :  8170 Mean   :11616 Mean   :1961
3rd Qu.:0.2500000 >Six  :  5358 >Six  :  5441 3rd Qu.:17407 3rd Qu.:1972
Max.   :0.2500000 Four   :  4120 Four   :  4150 Max.   :23268 Max.   :1995
              (Other): 2334 (Other): 3025

      dodm      doe      dox      doh      dod
Min.   :1933    Min.   :2005    Min.   :2007    Min.   :1985    Min.   :2007
1st Qu.:1978    1st Qu.:2007    1st Qu.:2012    1st Qu.:1998    1st Qu.:2010
Median :1989    Median :2007    Median :2012    Median :2003    Median :2011
Mean   :1987    Mean   :2008    Mean   :2012    Mean   :2003    Mean   :2011
3rd Qu.:1999    3rd Qu.:2008    3rd Qu.:2012    3rd Qu.:2007    3rd Qu.:2011
Max.   :2012    Max.   :2012    Max.   :2012    Max.   :2014    Max.   :2012
              NA's   :183691 NA's   :270319

      sex      bmi      hba      alb      sbp      aht
M:158642    Min.   :14.45    Min.   : 22.40    Min.   :  0.00    Min.   :  0.0    N:176558
F:120362    1st Qu.:22.60    1st Qu.: 55.19    1st Qu.:  0.50    1st Qu.:120.0    Y:102446
              Median :24.76    Median : 63.93    Median :  1.00    Median :130.0
              Mean   :25.35    Mean   : 65.80    Mean   :  7.01    Mean   :132.5
              3rd Qu.:27.45    3rd Qu.: 73.77    3rd Qu.:  2.30    3rd Qu.:141.0
              Max.   :62.87    Max.   :249.73    Max.   :1072.03    Max.   :250.0
              NA's   :107196

      llr      smk      gbmi      galb      nhh
N:186596    Never smoker :105660 Under-wt: 16113 Normo:164936 Min.   :0.0000
Y: 92408    Daily smoker  : 88326 Normal  :130633 Micro: 6298 1st Qu.:0.0000
              Ex smoker   : 36207 Over-wt : 99183 Macro: 574 Median :0.0000
              No information : 38130 Obese   : 33075 NA's :107196 Mean   :0.3416
              Occasional smoker: 10681 3rd Qu.:1.0000
              Max.   :1.0000

      anyHH      tfh      hh
Mode :logical    Min.   : 0.000    Mode :logical
FALSE:276738     1st Qu.: 0.000    FALSE:276738
TRUE :2266        Median : 0.000    TRUE :2266
NA's :0           Mean   : 1.724    NA's :0
              3rd Qu.: 1.255
              Max.   :26.618

```

6.1.0.1 Persons with at least one event If we want the number of persons that see at least one event during follow-up we must subset the Lh object to records with HH events:

```

> levels( Lh )
[1] "No HH" "One" "Two" "Three" "Four" "Five" ">Six" "Dead"

> Ls <- subset( Lh, !is.na(sbp) &
+               lex.Cst != lex.Xst &
+               lex.Xst != "Dead" )
> addmargins( table(table(Ls$lex.id)) )

      1      2      3      4      5      6 Sum
1271 267    77    34    14     4 1667

```

7 Modeling HH rates

First we provide the analysis dataset and an overview of it:

```
> Sh <- subset( Sh, !is.na(sbp) & lex.Cst!=">Six" )
> summary( Sh )
```

Transitions:

From	To	No HH	One	Two	Three	Four	Five	>Six	Dead	Records:	Events:	Risk time:
No HH		195274	851	0	0	0	0	0	315	196440	1166	47223.92
One		0	45458	521	0	0	0	0	138	46117	659	10998.20
Two		0	0	16114	345	0	0	0	60	16519	405	3891.96
Three		0	0	0	7825	240	0	0	51	8116	291	1890.37
Four		0	0	0	0	3910	176	0	34	4120	210	944.14
Five		0	0	0	0	0	2190	133	11	2334	144	527.63
Sum		195274	46309	16635	8170	4150	2366	133	609	273646	2875	65476.22

Transitions:

From	To	Persons:
No HH		13010
One		3413
Two		1421
Three		781
Four		444
Five		294
Sum		17230

```
> with( subset( Sh, lex.Xst != lex.Cst & lex.Xst != "Dead" ), table( lex.Cst, lex.Xst ) )
```

lex.Cst	lex.Xst	No HH	One	Two	Three	Four	Five	>Six	Dead
No HH		0	851	0	0	0	0	0	0
One		0	0	521	0	0	0	0	0
Two		0	0	0	345	0	0	0	0
Three		0	0	0	0	240	0	0	0
Four		0	0	0	0	0	176	0	0
Five		0	0	0	0	0	0	133	0
>Six		0	0	0	0	0	0	0	0
Dead		0	0	0	0	0	0	0	0

```
> ( tt <-
+ with( subset( Sh, lex.Xst != lex.Cst & lex.Xst != "Dead" ),
+       addmargins( table( table( lex.id ) ) ) ) )
```

	1	2	3	4	5	6	Sum
1271	267	77	34	14	4	4	1667

```
> sum( tt[1:6]*1:6 )
[1] 2266
```

```
> with( subset( Sh, lex.Xst != lex.Cst & lex.Xst != "Dead" ),
+       length(unique(lex.id)) )
[1] 1667
```

We model the effects of the time scales non-linearly using 4 knots (3 parameters) for each variable. We want the knots for the various effects to be placed so that we have the same number of events between each pair of knots (except for `tfh` — see below):

```
> eo <- subset( Sh, hh )
> ( kp <- with( eo, quantile( per+lex.dur, (1:4-0.5)/4 ) ) )
      12.5%      37.5%      62.5%      87.5%
2007.638 2009.002 2010.218 2011.424
```

```
> ( ka <- with( eo, quantile( age+lex.dur, (1:4-0.5)/4 ) ) )
```

```

      12.5%    37.5%    62.5%    87.5%
28.47365 44.29055 55.77139 71.49589
> ( kd <- with( eo, quantile( dur+lex.dur, (1:4-0.5)/4 ) ) )
      12.5%    37.5%    62.5%    87.5%
 8.180014 18.338809 29.004791 44.806982
> ( kh <- c( 0,
+           with( eo, quantile( tfh+lex.dur, 1:3 /3.5 ) ) ) )
      28.57143% 57.14286% 85.71429%
0.0000000 0.1500684 1.0157426 6.1179231

```

For reporting the effects of the time scales we need reference points on each of them, *except* the `tfh`-scale which is not defined for those in the “No HH” group.

Note that when defining knots for the timescale `tfh`, we explicitly included 0 as the first knot, meaning that the spline effects by definition are 0 at 0, and thus that *coding* `tfh` as 0 for the “No HH” level of `lex.Cst` (in which state “time since last HH” is not defined) means that the spline effects are exclusively estimated using information from the other levels of `lex.Cst`.

In the modeling we include all clinical variables except albumin measurement:

```

> m0 <- glm( hh ~ sex + I(bmi/5) + I(hba/10) + I(sbp/10) + smk + aht + llt +
+           Ns( per, knots=kp ) +
+           Ns( age, knots=ka ) +
+           Ns( dur, knots=kd ) +
+           Ns( tfh, knots=kh ) + lex.Cst,
+           offset = log( lex.dur ),
+           family = poisson,
+           data = Sh )
> mi <- update( m0, . ~ . - Ns( tfh, knots=kh )
+               + Ns( tfh, knots=kh ):lex.Cst )
> m1 <- update( m0, . ~ . - Ns( per, knots=kp )
+               + per )
> mli<- update( m1, . ~ . - per + I(per-2005):lex.Cst )
> anova( mli, m1, m0, mi, test="Chisq" )
Analysis of Deviance Table

```

```

Model 1: hh ~ sex + I(bmi/5) + I(hba/10) + I(sbp/10) + smk + aht + llt +
  Ns(age, knots = ka) + Ns(dur, knots = kd) + Ns(tfh, knots = kh) +
  lex.Cst + lex.Cst:I(per - 2005)
Model 2: hh ~ sex + I(bmi/5) + I(hba/10) + I(sbp/10) + smk + aht + llt +
  Ns(age, knots = ka) + Ns(dur, knots = kd) + Ns(tfh, knots = kh) +
  lex.Cst + per
Model 3: hh ~ sex + I(bmi/5) + I(hba/10) + I(sbp/10) + smk + aht + llt +
  Ns(per, knots = kp) + Ns(age, knots = ka) + Ns(dur, knots = kd) +
  Ns(tfh, knots = kh) + lex.Cst
Model 4: hh ~ sex + I(bmi/5) + I(hba/10) + I(sbp/10) + smk + aht + llt +
  Ns(per, knots = kp) + Ns(age, knots = ka) + Ns(dur, knots = kd) +
  lex.Cst + Ns(tfh, knots = kh):lex.Cst
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      273615      24080
2      273620      24088 -5   -8.5060  0.1305
3      273618      24088  2    0.3828  0.8258
4      273606      24073 12   15.0318  0.2397

```

Firstly, we see that there are no indication of interaction between times since HH and state of HH, the HRs are proportional; the formal p-value is 0.240. Moreover, there is neither any indication that the secular trend (change in the rates of HH by calendar time is non-linear); the p-value for linearity is 0.826, and finally there is no evidence that the linear trend is different between the states, and certainly no consistent pattern across the states:

```
> round( ci.exp( mli, subset="per", pval=TRUE ), 4 )
              exp(Est.)    2.5%    97.5%      P
lex.CstNo HH:I(per - 2005)    0.9079 0.8674 0.9502 0.0000
lex.CstOne:I(per - 2005)     0.9885 0.9319 1.0485 0.7002
lex.CstTwo:I(per - 2005)     0.8862 0.8253 0.9516 0.0009
lex.CstThree:I(per - 2005)    0.9182 0.8419 1.0014 0.0537
lex.CstFour:I(per - 2005)     0.9891 0.8934 1.0950 0.8326
lex.CstFive:I(per - 2005)     0.9296 0.8310 1.0398 0.2017
```

We can explore further non-linear effects separately for the clinical measurements:

```
> anova( ml, update( ml, . ~ . + I(bmi^2) ),
+       ml, update( ml, . ~ . + I(hba^2) ),
+       ml, update( ml, . ~ . + I(sbp^2) ), test="Chisq" )[1:3*2,]
Analysis of Deviance Table

Model 1: hh ~ sex + I(bmi/5) + I(hba/10) + I(sbp/10) + smk + aht + llt +
  Ns(age, knots = ka) + Ns(dur, knots = kd) + Ns(tfh, knots = kh) +
  lex.Cst + per
Model 2: hh ~ sex + I(bmi/5) + I(hba/10) + I(sbp/10) + smk + aht + llt +
  Ns(age, knots = ka) + Ns(dur, knots = kd) + Ns(tfh, knots = kh) +
  lex.Cst + per + I(bmi^2)
Model 3: hh ~ sex + I(bmi/5) + I(hba/10) + I(sbp/10) + smk + aht + llt +
  Ns(age, knots = ka) + Ns(dur, knots = kd) + Ns(tfh, knots = kh) +
  lex.Cst + per
Model 4: hh ~ sex + I(bmi/5) + I(hba/10) + I(sbp/10) + smk + aht + llt +
  Ns(age, knots = ka) + Ns(dur, knots = kd) + Ns(tfh, knots = kh) +
  lex.Cst + per + I(hba^2)
Model 5: hh ~ sex + I(bmi/5) + I(hba/10) + I(sbp/10) + smk + aht + llt +
  Ns(age, knots = ka) + Ns(dur, knots = kd) + Ns(tfh, knots = kh) +
  lex.Cst + per
Model 6: hh ~ sex + I(bmi/5) + I(hba/10) + I(sbp/10) + smk + aht + llt +
  Ns(age, knots = ka) + Ns(dur, knots = kd) + Ns(tfh, knots = kh) +
  lex.Cst + per + I(sbp^2)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
2    273619      24075  1  12.8795 0.0003322
4    273619      24088  1   0.3435 0.5578021
6    273619      24087  1   0.9916 0.3193508
```

Thus we see a strong non-linear effect of the HbA_{1c} (**hba**) measured at entry (**doe**), but no evidence of non-linearity of BMI or systolic blood pressure. Hence, we expand the model with a spline effect of **hba**:

```
> ( kH <- with( eo, quantile( hba, (1:4-0.5)/4 ) ) )
      12.5%    37.5%    62.5%    87.5%
50.81985 62.84175 72.67785 87.97845

> mx <- update( ml, . ~ . - I(hba/10) + Ns( hba, knots=kH ) )
> anova( ml, mx, test="Chisq" )
Analysis of Deviance Table

Model 1: hh ~ sex + I(bmi/5) + I(hba/10) + I(sbp/10) + smk + aht + llt +
  Ns(age, knots = ka) + Ns(dur, knots = kd) + Ns(tfh, knots = kh) +
  lex.Cst + per
Model 2: hh ~ sex + I(bmi/5) + I(sbp/10) + smk + aht + llt + Ns(age, knots = ka) +
  Ns(dur, knots = kd) + Ns(tfh, knots = kh) + lex.Cst + per +
  Ns(hba, knots = kH)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1    273620      24088
2    273618      24079  2   9.0411 0.01088
```

The model `mx` should be the model that we report in the paper, also the secular trend:

```
> cbind( round( ( 1 - ci.exp( mx, subset="per" ) ) * 100 , 1 ),
+         round( ci.exp( mx, subset="per", pval=TRUE), 4 ) )
      exp(Est.) 2.5% 97.5% exp(Est.) 2.5% 97.5% P
per          7  9.5  4.3    0.9304 0.9046 0.957 0
```

Thus the annual decrease in the rate of HH is 7.0% (95% CI 4.3;9.5), strongly significant.

Further, we extract the non-linear effects (HbA_{1c}, current age, DM duration and time since last HH) *relative* to the (arbitrarily) chosen reference points, using suitably chosen point to construct the relevant matrices. Note that there is no reference point for the effects of `tfh`, the estimated HRs will be relative the the follow up in the state “No HH”.

```
> summary.data.frame( Sh[,c("hba", "age", "per", "dur", "tfh")] )
      hba          age          per          dur          tfh
Min.   : 22.40   Min.   :16.06   Min.   :2005   Min.   : 0.00   Min.   : 0.0000
1st Qu.: 55.19   1st Qu.:36.92   1st Qu.:2008   1st Qu.:10.55   1st Qu.: 0.0000
Median : 63.93   Median :48.00   Median :2010   Median :19.77   Median : 0.0000
Mean   : 65.83   Mean   :48.23   Mean   :2010   Mean   :21.94   Mean   : 1.6526
3rd Qu.: 73.77   3rd Qu.:59.57   3rd Qu.:2011   3rd Qu.:31.55   3rd Qu.: 0.8209
Max.   :249.73   Max.   :98.60   Max.   :2012   Max.   :78.75   Max.   :26.6179

> rH <- 60
> rp <- 2010
> ra <- 50
> rd <- 20
> pH <- seq( 20, 120,,100) ; cH <- Ns( pH, knots=kH, ref=rH )
> pp <- seq(2005,2012,,100) ; cp <- Ns( pp, knots=kp, ref=rp )
> pa <- seq( 20, 80,,100) ; ca <- Ns( pa, knots=ka, ref=ra )
> pd <- seq( 0, 50,,100) ; cd <- Ns( pd, knots=kd, ref=rd )
> ph <- seq( 0, 15,,100) ; ch <- cbind( 1,
+                                       Ns( ph, knots=kh ) )
```

We now extract HbA_{1c} and the time-scale effects from the `mx` model

```
> RRH <- ci.exp( mx, subset="hba", ctr.mat=cH )
> RRa <- ci.exp( mx, subset="age", ctr.mat=ca )
> RRd <- ci.exp( mx, subset="dur", ctr.mat=cd )
> ci.exp( mx, subset=c("One", "tfh") )
      exp(Est.) 2.5% 97.5%
lex.CstOne    10.0296445 8.30182914 12.1170609
Ns(tfh, knots = kh)1 0.2774557 0.23266892 0.3308636
Ns(tfh, knots = kh)2 0.1261205 0.09124948 0.1743175
Ns(tfh, knots = kh)3 0.2555129 0.22308815 0.2926504

> RRh1 <- ci.exp( mx, subset=c("One", "tfh"), ctr.mat=ch )
> RRh2 <- ci.exp( mx, subset=c("Two", "tfh"), ctr.mat=ch )
> RRh3 <- ci.exp( mx, subset=c("Thr", "tfh"), ctr.mat=ch )
> RRh4 <- ci.exp( mx, subset=c("Fou", "tfh"), ctr.mat=ch )
> RRh5 <- ci.exp( mx, subset=c("Fiv", "tfh"), ctr.mat=ch )
```

These non-linear HR effects can then be plotted:

```
> par( mfrow=c(2,2), mar=c(3,3,1,2), mgp=c(3,1,0)/1.6, bty="n", las=1 )
> matplot( pH, RRH,
+          type="l", lty=1, col=1, lwd=c(3,1,1),
+          log="y", ylim=c(0.8,2),
+          xlab=expression("Hb"*A[1][c]*" (mmol/mol)"), ylab="HH hazard ratio" )
> abline( h=1 )
> rug( with( eo, quantile( hba, probs=0:10/10 ) ) )
> axis( side=1, at=seq(30,110,20), labels=NA, tcl=-0.3 )
> axis( side=2, at=9:19/10, labels=NA, tcl=-0.3 )
```

```

> text( 20, 2, "A", font=2, cex=1.3, adj=c(0,1) )
> matplot( pa, RRa,
+         type="l", lty=1, col=1, lwd=c(3,1,1),
+         log="y", ylim=c(0.8,2),
+         xlab="Age at follow-up (years)", ylab="HH hazard ratio" )
> abline( h=1 )
> rug( with( eo, quantile( age+lex.dur, probs=0:10/10 ) ) )
> axis( side=1, at=seq(25,75,10), labels=NA, tcl=-0.3 )
> axis( side=2, at=9:19/10, labels=NA, tcl=-0.3 )
> text( 20, 2, "B", font=2, cex=1.3, adj=c(0,1) )
> matplot( pd, RRd,
+         type="l", lty=1, col=1, lwd=c(3,1,1),
+         log="y", ylim=c(0.8,2),
+         xlab="DM duration at follow-up (years)", ylab="HH hazard ratio" )
> abline( h=1 )
> rug( with( eo, quantile( dur+lex.dur, probs=0:10/10 ) ) )
> axis( side=1, at=seq(5,45,10), labels=NA, tcl=-0.3 )
> axis( side=2, at=9:19/10, labels=NA, tcl=-0.3 )
> text( 0, 2, "C", font=2, cex=1.3, adj=c(0,1) )
> matplot( ph, cbind( RRh1, RRh2, RRh3, RRh4, RRh5 ),
+         type="l", lty=1, col=rep(gray(0:4/6),each=3), lwd=c(3,1,1),
+         log="y", ylim=c(0.9,25),
+         xlab="Time since last HH (years)", ylab="HH hazard ratio relative to "No HH" )
> abline( h=1 )
> rug( with( eo, quantile( tfh+lex.dur, probs=0:10/10 ) ) )
> axis( side=1, at=1:14, labels=NA, tcl=-0.3 )
> axis( side=4, at=c(1,2,5,10) )
> axis( side=4, at=c(0.9,1,1.5,2:10), labels=NA, tcl=-0.3 )
> axis( side=2, at=c(0.9,1,1.5,2:10,15), labels=NA, tcl=-0.3 )
> text( 1, 25, "D", font=2, cex=1.3, adj=c(0,1) )

```

Since the model for the effect of increasing number of HH is a proportional hazards model on the time since last HH time scale, we can report the constant HRs between successive states, that is 2 vs. 1, 3 vs. 2, ..., 5 vs. 4:

```

> prf <- c("Prev. HH:", " ") [c(1,2,2,2)]
> cdif <- contr.diff(5)[-1,]
> rownames(cdif) <- paste( prf, 2:5, "vs.", 1:4 )
> colnames(cdif) <- 1:5
> cdif
      1  2  3  4  5
Prev. HH: 2 vs. 1 -1  1  0  0  0
           3 vs. 2  0 -1  1  0  0
           4 vs. 3  0  0 -1  1  0
           5 vs. 4  0  0  0 -1  1

> mdif <- cbind(-1,diag(4))
> rownames(mdif) <- paste( prf, 2:5, "vs. 1" )
> colnames(mdif) <- 1:5
> mdif
      1  2  3  4  5
Prev. HH: 2 vs. 1 -1  1  0  0  0
           3 vs. 1 -1  0  1  0  0
           4 vs. 1 -1  0  0  1  0
           5 vs. 1 -1  0  0  0  1

> ( vsp <- ci.exp( mx, subset="Cst", ctr.mat=cdif, pval=T ) )
      exp(Est.)    2.5%    97.5%      P
Prev. HH: 2 vs. 1  1.588287 1.3833098 1.823638 5.290866e-11
           3 vs. 2  1.256115 1.0644127 1.482344 6.959575e-03
           4 vs. 3  1.317358 1.0838654 1.601151 5.622452e-03
           5 vs. 4  1.212425 0.9673571 1.519579 9.454283e-02

> ( vs1 <- ci.exp( mx, subset="Cst", ctr.mat=mdif, pval=T ) )

```

```

      exp(Est.)    2.5%    97.5%      P
Prev. HH: 2 vs. 1  1.588287 1.383310 1.823638 5.290866e-11
           3 vs. 1  1.995072 1.704062 2.335778 9.006678e-18
           4 vs. 1  2.628224 2.202646 3.136029 8.074270e-27
           5 vs. 1  3.186525 2.615564 3.882124 1.260659e-30

> hhr <- cbind( vsp, vs1 )
> rownames( hhr ) <- 2:5
> colnames( hhr )[c(1,5)] <- c("Vs. previous", "Vs. 1 HH")
> round( hhr, 2 )

  Vs. previous 2.5% 97.5%    P    Vs. 1 HH 2.5% 97.5% P
2          1.59 1.38  1.82 0.00          1.59 1.38  1.82 0
3          1.26 1.06  1.48 0.01          2.00 1.70  2.34 0
4          1.32 1.08  1.60 0.01          2.63 2.20  3.14 0
5          1.21 0.97  1.52 0.09          3.19 2.62  3.88 0

```

Thus we see that once beyond the first HH, it seems that the effects of each extra HH is either the same or slightly decreasing. We can devise a Wald test for the equality of these 4 differences by testing whether the second order differences are 0:

```

> ( c2dif <- cdif[-4,-5] %*% cdif )

      1 2 3 4 5
Prev. HH: 2 vs. 1 1 -2 1 0 0
           3 vs. 2 0 1 -2 1 0
           4 vs. 3 0 0 1 -2 1

> Wald( mx, subset="Cst", ctr.mat=c2dif )

      Chisq      d.f.      P
6.99598083 3.00000000 0.07202599

```

So formally there is no evidence that the HRs between the successive levels of HH are different, but it is not really convincing with a P-value of 7%. So the conclusion is that there is an *extra* effect of another HH on the occurrence on further HHs but also that this effect is diminishing by number previous of HHs.

Finally we can report the effect of the clinical variables on the occurrence of HH by extracting the parameters from the final model (`mx`):

```

> res <- rbind( ci.exp( mx, pval=TRUE )[c(2:10,25),], vs1, vsp )
> cbind( round( res[,-4], 2 ), round( res[,4,drop=F], 3 ) )

      exp(Est.) 2.5% 97.5%    P
sexF          0.89 0.82  0.97 0.011
I(bmi/5)       0.91 0.86  0.96 0.001
I(sbp/10)      1.00 0.98  1.03 0.850
smkDaily smoker 1.10 0.99  1.21 0.068
smkEx smoker   1.12 0.98  1.28 0.088
smkNo information 0.94 0.81  1.08 0.352
smkOccasional smoker 1.11 0.89  1.39 0.369
ahtY          1.24 1.12  1.37 0.000
lltY          1.06 0.96  1.17 0.253
per           0.93 0.90  0.96 0.000
Prev. HH: 2 vs. 1  1.59 1.38  1.82 0.000
           3 vs. 1  2.00 1.70  2.34 0.000
           4 vs. 1  2.63 2.20  3.14 0.000
           5 vs. 1  3.19 2.62  3.88 0.000
Prev. HH: 2 vs. 1  1.59 1.38  1.82 0.000
           3 vs. 2  1.26 1.06  1.48 0.007
           4 vs. 3  1.32 1.08  1.60 0.006
           5 vs. 4  1.21 0.97  1.52 0.095

> cbind( round( res[,-4], 2 ), round( res[,4,drop=F], 3 ) )[c(1:5,7,6,8:10,18:16,11:14),]

```

	exp(Est.)	2.5%	97.5%	P
sexF	0.89	0.82	0.97	0.011
I(bmi/5)	0.91	0.86	0.96	0.001
I(sbp/10)	1.00	0.98	1.03	0.850
smkDaily smoker	1.10	0.99	1.21	0.068
smkEx smoker	1.12	0.98	1.28	0.088
smkOccasional smoker	1.11	0.89	1.39	0.369
smkNo information	0.94	0.81	1.08	0.352
ahtY	1.24	1.12	1.37	0.000
lltY	1.06	0.96	1.17	0.253
per	0.93	0.90	0.96	0.000
5 vs. 4	1.21	0.97	1.52	0.095
4 vs. 3	1.32	1.08	1.60	0.006
3 vs. 2	1.26	1.06	1.48	0.007
Prev. HH: 2 vs. 1	1.59	1.38	1.82	0.000
3 vs. 1	2.00	1.70	2.34	0.000
4 vs. 1	2.63	2.20	3.14	0.000
5 vs. 1	3.19	2.62	3.88	0.000

Since the effect of previous HH depends strongly on time since last HH, we can only compute the effect of this with reference to specific times since HH. The median time of event is at 0.4 years after last HH, the 70th percentile is at 2.5 years, and the 90th at 8.1 years so we compute the HR associated with different numbers of previous HH (relative to no previous HH) for times 0.5, 1, 5 and 10 years:

```
> wht <- c(0.5,1,5,10)
> crh <- cbind( 1, Ns( wht, knots=kh ) )
> rlh <- NArray( list( "Time from HH" = wht,
+                      "HR vs. no HH" = c("HR", "lo", "hi"),
+                      "Prev. HHs" = 1:5 ) )
> rlh[, ,1] <- ci.exp( mx, subset=c("One", "tfh"), ctr.mat=crh )
> rlh[, ,2] <- ci.exp( mx, subset=c("Two", "tfh"), ctr.mat=crh )
> rlh[, ,3] <- ci.exp( mx, subset=c("Thr", "tfh"), ctr.mat=crh )
> rlh[, ,4] <- ci.exp( mx, subset=c("Fou", "tfh"), ctr.mat=crh )
> rlh[, ,5] <- ci.exp( mx, subset=c("Fiv", "tfh"), ctr.mat=crh )
> round( ftable( rlh, col.vars=1:2 ), 1 )
```

	Time from HH	0.5		1		5		10					
	HR vs. no HH	HR	lo	hi	HR	lo	hi	HR	lo	hi	HR	lo	hi
Prev. HHs													
1		5.5	4.8	6.5	4.3	3.7	4.9	2.0	1.8	2.3	1.7	1.5	2.0
2		8.8	7.5	10.4	6.8	5.8	7.9	3.3	2.8	3.8	2.7	2.3	3.2
3		11.1	9.3	13.2	8.5	7.2	10.0	4.1	3.5	4.8	3.4	2.9	4.2
4		14.6	12.1	17.6	11.2	9.3	13.4	5.4	4.5	6.5	4.5	3.7	5.6
5		17.7	14.4	21.7	13.6	11.1	16.6	6.5	5.3	8.0	5.5	4.4	6.9

Note that these HRs are merely 4 select points on the 5 curves in the lower right hand panel of figure 7; and also that the successive hazard ratios between states is the same at all times (the proportional hazards assumption).

7.0.1 Occurrence rates for 1st HH

The model also allows us to predict absolute occurrence rates of HH for any combination of covariates, and since age has the strongest non-linear pattern we choose to predict rates as a function of age using values for other variables close to the median:

```
> a.pr <- 20:80
> nd <- data.frame( age = c(a.pr, NA, a.pr),
+                  sex = "M",
+                  bmi = 25,
```

```

+             hba = 65,
+             sbp = 130,
+             aht = "N",
+             llt = "N",
+             smk = "Never smoker",
+             dur = 20,
+             tfh = 0,
+             lex.Cst = "No HH",
+             per = c( rep(2006,length(a.pr)),NA,
+                     rep(2012,length(a.pr)) ),
+             lex.dur = 100 )
> str( nd )

'data.frame':      123 obs. of  13 variables:
 $ age      : int   20 21 22 23 24 25 26 27 28 29 ...
 $ sex      : Factor w/ 1 level "M": 1 1 1 1 1 1 1 1 1 1 ...
 $ bmi      : num   25 25 25 25 25 25 25 25 25 25 ...
 $ hba      : num   65 65 65 65 65 65 65 65 65 65 ...
 $ sbp      : num  130 130 130 130 130 130 130 130 130 130 ...
 $ aht      : Factor w/ 1 level "N": 1 1 1 1 1 1 1 1 1 1 ...
 $ llt      : Factor w/ 1 level "N": 1 1 1 1 1 1 1 1 1 1 ...
 $ smk      : Factor w/ 1 level "Never smoker": 1 1 1 1 1 1 1 1 1 1 ...
 $ dur      : num   20 20 20 20 20 20 20 20 20 20 ...
 $ tfh      : num    0 0 0 0 0 0 0 0 0 0 ...
 $ lex.Cst  : Factor w/ 1 level "No HH": 1 1 1 1 1 1 1 1 1 1 ...
 $ per      : num  2006 2006 2006 2006 2006 ...
 $ lex.dur  : num   100 100 100 100 100 100 100 100 100 100 ...

> prx <- ci.pred( mx, nd )
> matplot( nd$age, prx,
+          type="l", lty=1, col=1, lwd=c(3,1,1),
+          log="y", ylim=c(0.8,4),
+          xlab="Age at follow-up (years)",
+          ylab="Incidence rate of first HH per 100 PY" )
> axis( side=1, at=seq(25,75,5), labels=NA, tcl=-0.3 )
> axis( side=2, at=c(8:19,25,35,45)/10, labels=NA, tcl=-0.3 )
> text( 20, prx[1,1]*1.02, "2006", font=1, adj=c(0,0) )
> text( 70, prx[113,1]/1.05, "2012", font=1, adj=c(0,0) )

```

8 Sensitivity modeling

We can explore the the robustness of the modeling by either *excluding* systolic blood pressure (sbp) from the models, and thereby including some more patients, or *including* albumin levels (alb) thereby excluding further patients.

So first we provide an overview of the data available for the different analyses:

```

> npitt <- c(
+ with( mx$data,
+       length( unique( lex.id ) ) ),
+ with( subset( mx$data, !is.na(sbp) ),
+       length( unique( lex.id ) ) ),
+ with( subset( mx$data, !is.na(alb) & !is.na(sbp) ),
+       length( unique( lex.id ) ) ),
+ with( subset( mx$data, alb>0 & !is.na(alb) & !is.na(sbp) ),
+       length( unique( lex.id ) ) ) )
> names( npitt ) <- c("All", "sbp", "sbp & alb", "sbp & alb>0")
> npitt

      All      sbp  sbp & alb sbp & alb>0
17230  17230   10916    9952

```

Thus the exclusion of the `sbp` from the analyses are not likely to change much, but inclusion of `alb` will bring the number of available persons substantially down, so the latter is the only one we bother to do.

8.1 Including albumin in the model

Albumin levels should be included in the model, transformed by $\log_{10}$, giving the effect of a 10-fold increase in albumin, corresponding to the difference in the thresholds for micro- and macro-albuminuria:

```
> mxa <- update( mx, ~ . + log10(ifelse(alb==0,NA,alb)) )

> esl <- cbind( ci.exp( mxa, pval=TRUE )[c(2:10,25,29),],
+             rbind( ci.exp( mx , pval=TRUE )[c(2:10,25),], NA ) ) [c(10,1,4,7,5,6,2:3,11,8,9),]
> newlab <- c("HR/cal.year",
+            "F vs. M",
+            "Smoker vs. no",
+            "Occ. vs. no",
+            "Ex sm. vs. no",
+            "Unkn. vs. no",
+            "BMI (5 kg/m2)",
+            "Sys BP (10 mmHg)",
+            "Albumin (10 fold)",
+            "Antihypertensive",
+            "Lipid lowering")
> cbind( rownames( esl ), newlab )

      [1,] "per"                newlab
      [2,] "sexF"               "HR/cal.year"
      [3,] "smkDaily smoker"    "F vs. M"
      [4,] "smkOccasional smoker" "Smoker vs. no"
      [5,] "smkEx smoker"       "Occ. vs. no"
      [6,] "smkNo information"   "Ex sm. vs. no"
      [7,] "I(bmi/5)"           "Unkn. vs. no"
      [8,] "I(sbp/10)"          "BMI (5 kg/m2)"
      [9,] "log10(ifelse(alb == 0, NA, alb))" "Sys BP (10 mmHg)"
     [10,] "ahtY"               "Albumin (10 fold)"
     [11,] "lltY"               "Antihypertensive"
                                "Lipid lowering"

> rownames( esl ) <- newlab
```

We can then print the table of effects from two different models, the first chunk of rows are from the model restricted to those with available albumin measurements, the rightmost chunk are estimates from the part of the dataset where only those with missing systolic blood pressure were excluded.

```
> round( esl, 3 )
```

	exp(Est.)	2.5%	97.5%	P	exp(Est.)	2.5%	97.5%	P
HR/cal.year	0.916	0.883	0.951	0.000	0.930	0.905	0.957	0.000
F vs. M	0.916	0.818	1.026	0.131	0.895	0.821	0.975	0.011
Smoker vs. no	1.024	0.899	1.166	0.726	1.098	0.993	1.213	0.068
Occ. vs. no	0.902	0.683	1.192	0.470	1.108	0.886	1.387	0.369
Ex sm. vs. no	0.983	0.828	1.167	0.847	1.122	0.983	1.281	0.088
Unkn. vs. no	0.760	0.621	0.929	0.007	0.936	0.815	1.076	0.352
BMI (5 kg/m2)	0.959	0.895	1.027	0.235	0.911	0.863	0.962	0.001
Sys BP (10 mmHg)	0.970	0.937	1.004	0.081	1.002	0.978	1.027	0.850
Albumin (10 fold)	1.279	1.172	1.396	0.000	NA	NA	NA	NA
Antihypertensive	1.066	0.928	1.224	0.367	1.236	1.116	1.369	0.000
Lipid lowering	1.020	0.894	1.163	0.770	1.060	0.959	1.171	0.253

We can compare these two sets of estimates in a forest plot:

```
> par( mar=c(3,3,1,1), mgp=c(3,1,0)/1.6 )
> plotEst( esl[,1:3], y=11:1-0.1, textpos=11:1, col="grey" , lwd=3,
+         cex=1.5, vref=1, xtic=6:15/10, grid=c(5:9/10,1,1.2,1.5),
+         xlog=TRUE, xlab="HH hazard ratio" )
> linesEst( esl[,5:7], y=11:1+0.1, textpos=11:1, col="black", lwd=3, cex=1.5 )

> eslx      <- esl[,5:7]
> eslx[9,]  <- esl[9,1:3]
> plotEst( eslx, y=11:1, textpos=11:1, col="black", lwd=3,
+         cex=1.5, vref=1, xtic=6:15/10, grid=c(5:9/10,1,1.2,1.5),
+         xlog=TRUE, xlab="HH hazard ratio" )
```

From figure 9 it is clear that there are no major differences between the estimates from the large dataset and the smaller dataset resulting from restricting analysis to persons with known measurements of albumin > 0 . This the primary reporting of the clinical effects on the incidence rates of HH will be the estimates based on the model *including* SBP, but *excluding* Albumin levels. The effect of albumin will be reported separately with a note that it is based on a slightly different model.

8.2 Categorizing variables

Although it is in general bad practice we fit a variant of the model `mx` where the linear effect of BMI and albumin are replaced by the usual categorizations. First we add the categorized versions of the variables to the model and see that they add little:

```
> mxG <- update( mxa, . ~ . + Relevel( gbmi, 2 ) + galb )
> anova( mxG, mxa, test="Chisq" )
Analysis of Deviance Table

Model 1: hh ~ sex + I(bmi/5) + I(sbp/10) + smk + aht + llt + Ns(age, knots = ka) +
  Ns(dur, knots = kd) + Ns(tfh, knots = kh) + lex.Cst + per +
  Ns(hba, knots = kH) + log10(ifelse(alb == 0, NA, alb)) +
  Relevel(gbmi, 2) + galb
Model 2: hh ~ sex + I(bmi/5) + I(sbp/10) + smk + aht + llt + Ns(age, knots = ka) +
  Ns(dur, knots = kd) + Ns(tfh, knots = kh) + lex.Cst + per +
  Ns(hba, knots = kH) + log10(ifelse(alb == 0, NA, alb))
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      153501      13771
2      153506      13775 -5   -3.7873   0.5804

> round( rbind( ci.exp( mxa, subset="bmi", pval=TRUE ),
+               ci.exp( mxG, subset="bmi", pval=TRUE ) ), 3 )

               exp(Est.)  2.5% 97.5%    P
I(bmi/5)          0.959 0.895 1.027 0.235
I(bmi/5)          1.091 0.938 1.269 0.258
Relevel(gbmi, 2)Under-wt  1.144 0.899 1.455 0.273
Relevel(gbmi, 2)Over-wt   0.891 0.747 1.063 0.201
Relevel(gbmi, 2)Obese     0.704 0.479 1.035 0.075

> round( rbind( ci.exp( mxa, subset="alb", pval=TRUE ),
+               ci.exp( mxG, subset="alb", pval=TRUE ) ), 3 )

               exp(Est.)  2.5% 97.5%    P
log10(ifelse(alb == 0, NA, alb))  1.279 1.172 1.396 0.000
log10(ifelse(alb == 0, NA, alb))  1.243 1.097 1.410 0.001
galbMicro                     1.078 0.814 1.426 0.601
galbMacro                     1.200 0.679 2.120 0.530
```

We finally fit the model where BMI and albumin have been replaced by the grouped versions only, using normal weight ($\text{BMI} \in [20; 25)$) and Normo-albuminuria as references

```
> mxg <- update( mx, . ~ . - I(bmi/5) + Relevel( gbmi, 2 ) + galb )
> round( ci.exp( mxg, subset=c("gb","ga"), pval=TRUE ), 3 )
```

	exp(Est.)	2.5%	97.5%	P
Relevel(gbmi, 2)Under-wt	1.129	0.924	1.381	0.235
Relevel(gbmi, 2)Over-wt	0.942	0.837	1.060	0.322
Relevel(gbmi, 2)Obese	0.864	0.717	1.041	0.125
galbMicro	1.494	1.217	1.835	0.000
galbMacro	1.975	1.211	3.221	0.006

So in this version we see the same decrease in HH rates by the two variables, but here in a non-quantifiable way.

9 Overview of baseline data

The main analysis is based on persons for whom the systolic blood pressure (**sbp**) is available, so here we construct the baseline table for all relevant variables based on this dataset;

```
> summary( L0 )
```

Transitions:

From	To	No HH	Dead	Records:	Events:	Risk time:	Persons:
No HH	No HH	17612	673	18285	673	70002.17	18285

```
> summary( Lh )
```

Transitions:

From	To	No HH	One	Two	Three	Four	Five	>Six	Dead	Records:	Events:	Risk time:	Persons:
No HH	No HH	12425	882	0	0	0	0	0	320	13627	1202	49663.18	13627
One	One	0	2860	549	0	0	0	0	143	3552	692	11449.02	3552
Two	Two	0	0	1063	367	0	0	0	60	1490	427	4089.07	1490
Three	Three	0	0	0	514	251	0	0	54	819	305	1972.07	819
Four	Four	0	0	0	0	243	182	0	34	459	216	976.01	459
Five	Five	0	0	0	0	0	154	138	11	303	149	542.81	303
>Six	>Six	0	0	0	0	0	0	353	51	404	51	1310.00	404
Sum	Sum	12425	3742	1612	881	494	336	491	673	20654	3042	70002.17	18285

```
> Lb <- subset( Lh, !duplicated(lex.id) & !is.na(sbp) )
> Lb <- subset( Lb, lex.Cst != ">Six" )
> summary( Lb )
```

Transitions:

From	To	No HH	One	Two	Three	Four	Five	>Six	Dead	Records:	Events:	Risk time:	Persons:
No HH	No HH	11844	851	0	0	0	0	0	315	13010	1166	47223.92	13010
One	One	0	2101	367	0	0	0	0	94	2562	461	9257.56	2562
Two	Two	0	0	655	210	0	0	0	35	900	245	3002.27	900
Three	Three	0	0	0	294	118	0	0	24	436	142	1321.60	436
Four	Four	0	0	0	0	117	73	0	14	204	87	601.46	204
Five	Five	0	0	0	0	0	66	48	4	118	52	304.13	118
Sum	Sum	11844	2952	1022	504	235	139	48	486	17230	2153	61710.94	17230

```
> # Just to check that we get the same numer of people
> # (but of course not PY and transitions):
> summary( subset( Sh, !duplicated(lex.id) ) )
```

Transitions:

From	To	No	HH	One	Two	Three	Four	Five	>Six	Dead	Records:	Events:	Risk time:	Persons:
No	HH	12988	19	0	0	0	0	0	0	3	13010	22	1509.34	13010
One		0	2541	20	0	0	0	0	0	1	2562	21	295.98	2562
Two		0	0	887	13	0	0	0	0	0	900	13	100.27	900
Three		0	0	0	431	5	0	0	0	0	436	5	48.29	436
Four		0	0	0	0	201	3	0	0	0	204	3	23.52	204
Five		0	0	0	0	0	117	1	0	0	118	1	13.49	118
Sum		12988	2560	907	444	206	120	1	4		17230	65	1990.89	17230

Thus we can devise IQR for the relevant quantitative variables for the persons entering the analyses, and tables of persons in different categories of the categorical variables:

```
> names( Lb )
[1] "per"      "age"      "dur"      "hh1"      "hh2"      "hh3"      "hh4"      "hh5"
[9] "hh6"      "lex.dur"  "lex.Cst"  "lex.Xst"  "lex.id"   "id"       "dob"      "dodm"
[17] "doe"      "dox"      "doh"      "dod"      "sex"      "bmi"      "hba"      "alb"
[25] "sbp"      "aht"      "llt"      "smk"      "gbmi"     "galb"     "nhh"

> ( cont.var <- names( Lb )[c(17,10,2:3,23,25,22,24)] )
[1] "doe"      "lex.dur"  "age"      "dur"      "hba"      "sbp"      "bmi"      "alb"

> cbind( cont.var, nnam <- c("Date",
+                             "PY",
+                             "Age",
+                             "DM duration",
+                             "HbA1c",
+                             "SysBP",
+                             "BMI",
+                             "Albumin") )

      cont.var
[1,] "doe"      "Date"
[2,] "lex.dur"  "PY"
[3,] "age"      "Age"
[4,] "dur"      "DM duration"
[5,] "hba"      "HbA1c"
[6,] "sbp"      "SysBP"
[7,] "bmi"      "BMI"
[8,] "alb"      "Albumin"

> iqr <- NArray( list( var = cont.var,
+                       iqr = paste(pct<-c(1,5,10,15,19)*5,"%") ) )
> for( vn in cont.var ) iqr[vn,] <- quantile( Lb[,vn], pct/100, na.rm=TRUE )
> rownames( iqr ) <- nnam
> ( cat.var <- names( Lb )[cn<-c(21,11,26:30)] )
[1] "sex"      "lex.Cst"  "aht"      "llt"      "smk"      "gbmi"     "galb"

> cbind( cat.var, cnam <- c("Sex",
+                             "No. prev. HH",
+                             "Antihypertensive",
+                             "Lipid lowering",
+                             "Smoking",
+                             "BMI level",
+                             "Albuminuria") )

      cat.var
[1,] "sex"      "Sex"
[2,] "lex.Cst"  "No. prev. HH"
[3,] "aht"      "Antihypertensive"
[4,] "llt"      "Lipid lowering"
[5,] "smk"      "Smoking"
[6,] "gbmi"     "BMI level"
[7,] "galb"     "Albuminuria"
```

```

> vtab <- NULL
> for( vn in 1:length(cn) )
+ {
+   vt <- cbind( tt <- table(Lb[,cn[vn]],useNA="ifany"), tt/sum(tt)*100 )
+   llg <- nchar( cnam[vn] )
+   rownames( vt ) <- paste( rep( c(cnam[vn],paste(rep(" ",llg),collapse="")),
+                               c(1,nrow(vt)-1) ),
+                             rownames(vt) )
+   vtab <- rbind( vtab, vt )
+ }
> vtab <- rbind( Total=apply(vtab[1:2,],2,sum), vtab )
> colnames( vtab ) <- c("N","%" )
> vtab <- vtab[vtab[,1]>0,]
> format( as.Date.cal.yr( iqr[1,] ), "%d %b %Y" )

      5 %      25 %      50 %      75 %      95 %
"15 Mar 2006" "01 Feb 2007" "15 Jun 2007" "15 Feb 2009" "28 Feb 2011"

> format( as.Date.cal.yr( iqr[1,] ), "%d/%m/%Y" )

      5 %      25 %      50 %      75 %      95 %
"15/03/2006" "01/02/2007" "15/06/2007" "15/02/2009" "28/02/2011"

> round( iqr, 2 )

var      iqr      5 %      25 %      50 %      75 %      95 %
Date      2006.20 2007.08 2007.45 2009.12 2011.16
PY         0.57   2.24   4.11   4.84   5.67
Age        20.00  34.08  45.60  57.67  71.70
DM duration 0.71   7.16  16.28  28.25  45.58
HbA1c      44.00  55.19  63.93  73.77  95.63
SysBP      108.00 120.00 130.00 141.00 165.00
BMI        19.72  22.55  24.77  27.47  33.14
Albumin    0.00   0.50   1.00   2.26   22.82

> round( vtab, 1 )

      N      %
Total 17230 100.0
Sex M  9873  57.3
     F  7357  42.7
No. prev. HH No HH 13010 75.5
      One  2562  14.9
      Two   900   5.2
      Three  436   2.5
      Four  204   1.2
      Five  118   0.7
Antihypertensive N 10968 63.7
                  Y  6262 36.3
Lipid lowering N 11464 66.5
                 Y  5766 33.5
Smoking Never smoker 6808 39.5
        Daily smoker 5248 30.5
        Ex smoker    2313 13.4
        No information 2232 13.0
        Occasional smoker 629 3.7
BMI level Under-wt 1051 6.1
         Normal    7967 46.2
         Over-wt   6035 35.0
         Obese     2177 12.6
Albuminuria Normo 10475 60.8
             Micro   400  2.3
             Macro   41   0.2
             NA      6314 36.6

```

These are the numbers that should be in the baseline table, table 1.

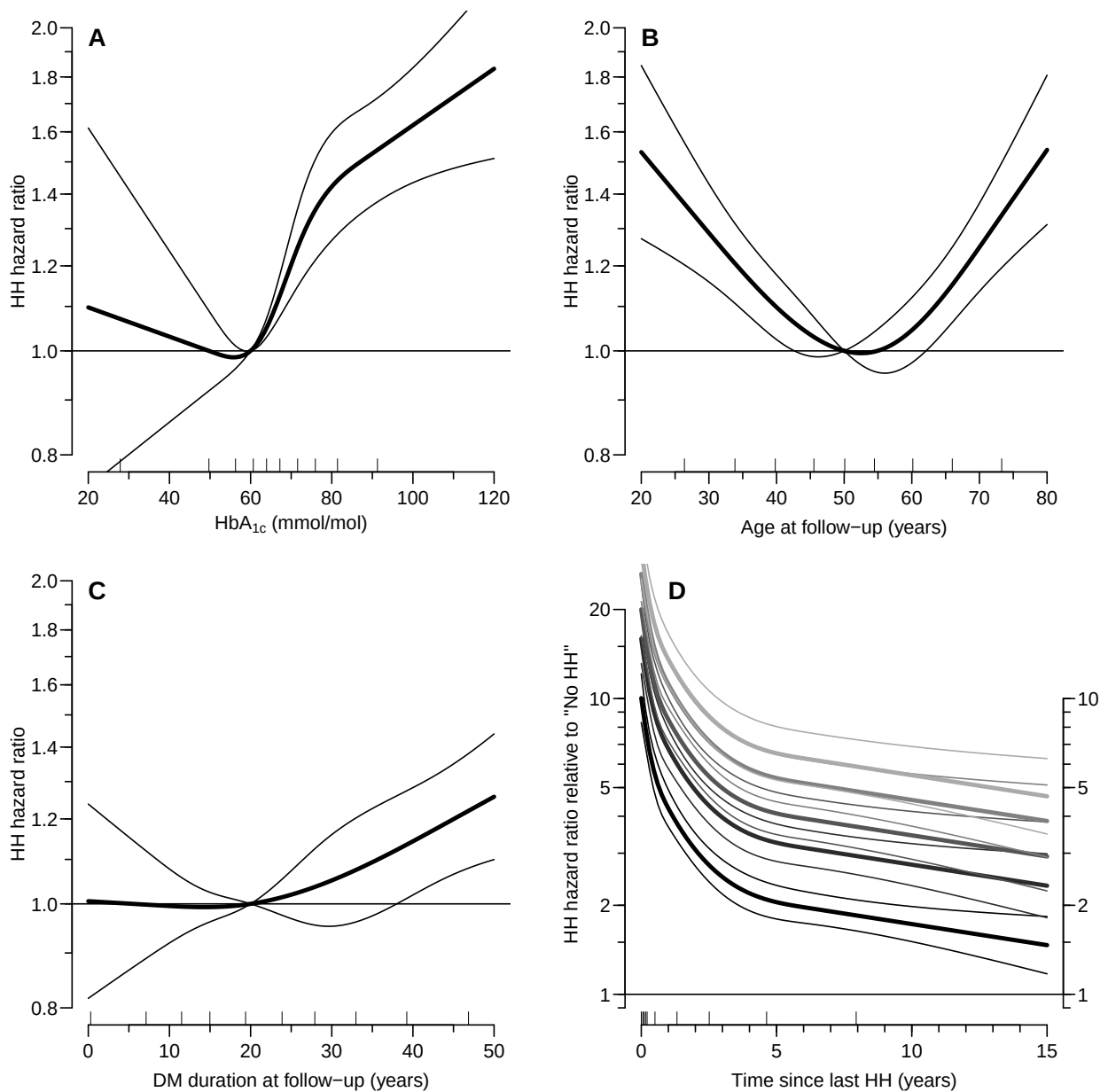


Figure 7: *Estimated time-effects on HH. The broken lines are estimates from the interaction model. They are hardly visible for the calendar time, age and duration plots, because the two sets of estimates are practically equal.*

In the upper left plot is also shown the estimated constant decrease in HH rates by calendar time (the straight line).

The lower right plot shows the HR relative to the "No HH" level, one curve for each of the levels 1, 2, 3 and 4 previous HHs; these are increasing.

The rug on the upside of the x-axis indicate the deciles of the HH events by each of the variables.

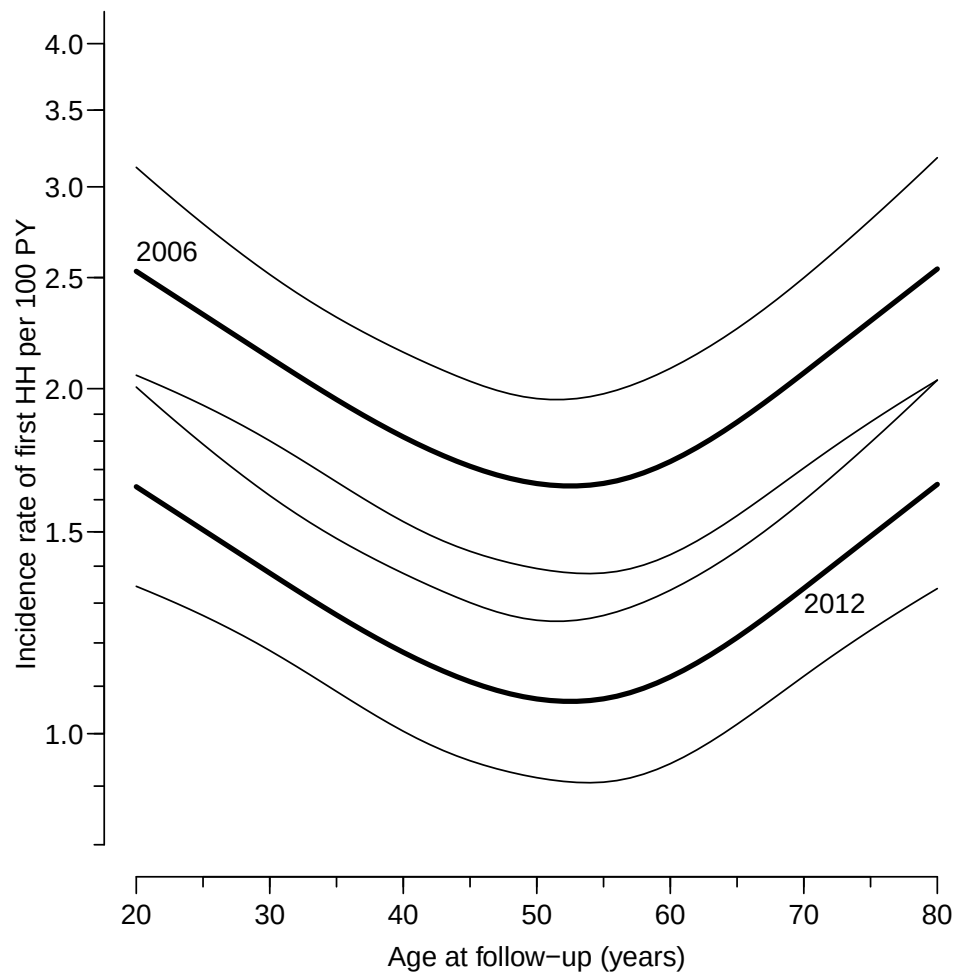


Figure 8: Predicted incidence rates of first HH occurrence by age at 1 January 2006 and 2012 (start and end of study), respectively. Prediction is for a never smoking man with 20 years of T1D duration, not on anti-hypertensive or lipid lowering medication, with a BMI of 25 kg/m^2 , systolic blood pressure 130 mmHg and a HbA_{1c} of 65 mmol/mol.

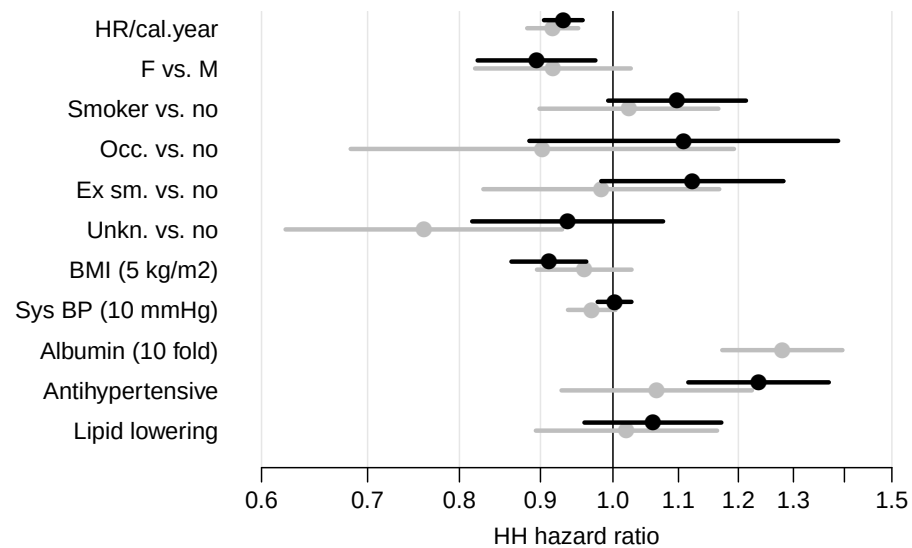


Figure 9: *Estimates of effects of clinical variables from the model without albumin level in black, and from the model including albumin level in gray.*