Diabetes and Tuberculosis in Denmark

SDC
June 2013
http://BendixCarstensen.com/DMreg/DMTB
Version 2.2

Compiled Tuesday 7th April, 2015, 14:48
from: /home/bendix/sdc/coll/zzkj/r/dmtb.tex

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

Contents

Chapter 1

Reading and setting up follow-up data

1.1 Data conversion

First we convert the data from SAS format to xport format which is R-readable:

```
1                                    "Program: getdata.sas"     17:15 Wednesday, October 24, 2012

NOTE: Copyright (c) 2002-2008 by SAS Institute Inc., Cary, NC, USA.
NOTE: SAS (r) Proprietary Software 9.2 (TS2M3)
      Licensed to NOVO NORDISK - BASIC PACKAGE, Site 50800704.
NOTE: This session is executing on the W32_VSPRO  platform.



NOTE: SAS initialization used:
      real time           3.79 seconds
      cpu time            0.59 seconds


NOTE: AUTOEXEC processing beginning; file is c:\stat\sas\autoexec.sas.

---------------------------------------------------------------
C:\Bendix\Steno\MaEJ\Tub-DM\sas\getdata.sas
---------------------------------------------------------------
NOTE: Libref HER was successfully assigned as follows:
      Engine:        V9
      Physical Name: C:\Bendix\Steno\MaEJ\Tub-DM\sas
NOTE: Libref DATA was successfully assigned as follows:
      Engine:        V9
      Physical Name: C:\Bendix\Steno\MaEJ\Tub-DM\data

NOTE: AUTOEXEC processing completed.

1          libname source 'P:\MAEJ\SAS data\DM&TB' ;
NOTE: Libref SOURCE was successfully assigned as follows:
      Engine:        V9
      Physical Name: P:\MAEJ\SAS data\DM&TB
2
3          data dmtb ;
4            set source.dmtb ;
5            drop V_PNR ;
6          run ;

WARNING: The variable V_PNR in the DROP, KEEP, or RENAME list has never been referenced.
NOTE: There were 1068322 observations read from the data set SOURCE.DMTB.
NOTE: The data set WORK.DMTB has 1068322 observations and 8 variables.
NOTE: DATA statement used (Total process time):
      real time           11.73 seconds
      cpu time            0.76 seconds


7
8          proc contents  data = dmtb ;
9              run ;

NOTE: PROCEDURE CONTENTS used (Total process time):
      real time           0.15 seconds
      cpu time            0.07 seconds

NOTE: The PROCEDURE CONTENTS printed page 1.

10
11         proc print  data = dmtb (obs=50) ;
12             run ;
```

```
NOTE: There were 50 observations read from the data set WORK.DMTB.
NOTE: The PROCEDURE PRINT printed page 2.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds


13
14         libname xptout xport '../data/dmtb.xpt' ;
NOTE: Libref XPTOUT was successfully assigned as follows:
      Engine:        XPORT
      Physical Name: C:\Bendix\Steno\MaEJ\Tub-DM\data\dmtb.xpt
15         proc copy  in = work  out = xptout  memtype = data ;
16            select dmtb ;
17         run;

NOTE: Copying WORK.DMTB to XPTOUT.DMTB (memtype=DATA).
NOTE: There were 1068322 observations read from the data set WORK.DMTB.
NOTE: The data set XPTOUT.DMTB has 1068322 observations and 8 variables.
NOTE: PROCEDURE COPY used (Total process time):
      real time           3.15 seconds
      cpu time            0.68 seconds


NOTE: SAS Institute Inc., SAS Campus Drive, Cary, NC USA 27513-2414
NOTE: The SAS System used:
      real time           19.09 seconds
      cpu time            2.16 seconds

The SAS System                                                          17:15 Wednesday, October 24, 2012   1

The CONTENTS Procedure

Data Set Name        WORK.DMTB                            Observations          1068322
Member Type          DATA                                 Variables             8
Engine               V9                                   Indexes               0
Created              24. oktober 2012 onsdag 17:15:42    Observation Length    72
Last Modified        24. oktober 2012 onsdag 17:15:42    Deleted Observations  0
Protection                                                Compressed            NO
Data Set Type                                             Sorted                NO
Label
Data Representation  WINDOWS_32
Encoding             wlatin1  Western (Windows)


                                  Engine/Host Dependent Information

Data Set Page Size          8192
Number of Data Set Pages    9455
First Data Page             1
Max Obs per Page            113
Obs in First Data Page      88
Number of Data Set Repairs  0
Filename                    C:\Users\BXC\AppData\Local\Temp\SAS Temporary Files\_TD6772\dmtb.sas7bdat
Release Created             9.0202M3
Host Created                W32_VSPRO


Alphabetic List of Variables and Attributes

#    Variable    Type    Len

8    doBTH       Num       8
1    doDM        Num       8
2    doDTH       Num       8
5    doIND       Num       8
3    doTB        Num       8
6    doUD        Num       8
4    region      Char      9
7    sex         Num       8

The SAS System                                                          17:15 Wednesday, October 24, 2012   2

    Obs     doDM    doDTH     doTB    region       doIND     doUD     sex     doBTH

      1    10981    11499       .                      .        .      2     -21914
      2    16700        .       .                      .        .      2      14610
      3        .        .       .    Europe       15183        .      1      14610
      4        .        .       .    Asia         15369        .      1      14610
      5        .        .       .    Asia         15519        .      1      14610
      6        .        .       .    Asia         16212        .      1      14610
      7        .        .       .    Asia         16349        .      1      14610
      8        .        .       .    Europe       14809        .      2      14610
      9        .        .       .    Asia         18192        .      1      14610
     10        .        .       .    Africa       18225        .      1      14610
     11        .        .       .    Asia         15018        .      2      14610
     12        .        .       .    Europe       15157        .      2      14610
     13        .        .       .    Asia         15635        .      2      14610
```

```
14        .        .        .    Asia          15910        .     2      14610
15        .        .        .    Africa        15937        .     2      14610
16    17757        .        .    Africa        15937        .     2      14610
17        .        .        .    Africa        16027        .     2      14610
18        .        .        .    Asia          16783        .     2      14610
19        .        .        .    Asia          17504    18058     2      14610
20        .        .        .    Africa        17015        .     2      14610
21        .        .        .    Asia          18074        .     2      14610
22        .        .        .    Asia          18240        .     2      14610
23    12227    12460        .                      .        .     2     -21549
24        .        .        .    Asia          15463        .     2      14976
25        .        .        .    Europe        16439        .     2      14976
26        .        .        .    Africa        17015        .     2      14976
27        .        .        .    Africa        17570        .     2      14976
28        .        .        .    Asia          17759        .     2      14976
29        .        .        .    Oceania       15142    15192     1      14976
30        .        .        .    Europe        15294        .     1      14976
31        .        .        .    America       15328        .     1      14976
32        .        .        .    Europe        15521        .     1      14976
33        .        .        .    Africa        16827        .     1      14976
34        .        .        .    Africa        17570        .     1      14976
35        .        .        .    Africa        18231        .     1      14976
36        .        .    18648                      .        .     1      14976
37    16713        .        .                      .        .     1      15341
38        .        .        .    Other         15503    15843     1      15341
39        .        .        .    America       15510        .     1      15341
40        .        .        .    Africa        16400        .     1      15341
41        .        .        .    Oceania       16596        .     1      15341
42        .        .        .    Africa        16622        .     1      15341
43        .        .        .    Asia          15516        .     2      15341
44        .        .        .    Africa        15937        .     2      15341
45        .        .        .    Europe        17512        .     2      15341
46        .        .        .    Asia          17191        .     1      15341
47        .        .        .    Europe        17328        .     1      15341
48        .        .        .    East_Euro     17387    17591     1      15341
49        .        .        .    East_Euro     18057    18473     1      15341
50        .        .        .    Africa        18225        .     1      15341
```

1.2 Data entry

The data in the just created SAS-xport file which contains records of all person who either

- have non-Danish born parents or
- a diagnosis of TB or
- a diagnosis of DM

Thus the only persons not included here are persons with Danish born parents and no record of either DM or TB.

First we read data and then groom the dataset a little:

```
> options( width=95 )
> memory.size(3500)
[1] 3500
```

```
> library(Epi)
> library(foreign)
> dmtb <- read.xport("../data/dmtb.xpt")
> names( dmtb ) <- tolower( names(dmtb) )
```

Sanity check: `region="" & DM FALSE & TB FALSE` should be 0:

```
> with( dmtb, ftable( region, Dead=!is.na(dodth),
+                             DM=!is.na(dodm),
+                             TB=!is.na(dotb), col.vars=4:2 ) )
          TB    FALSE                        TRUE
          DM    FALSE         TRUE         FALSE          TRUE
          Dead  FALSE   TRUE  FALSE   TRUE  FALSE   TRUE  FALSE   TRUE
region
```

```
                   68  12814 258154 162034   3479    156    219    148
Africa          31431    409   1262     87   1211     43     71      4
America         47138    275    380     28     26      2      0      0
Asia           122446   1443   5638    522    818     29     74     13
East_Euro      118367   1590   2537    449    172     13     14      4
Europe         243786   2896   4604    517    373     23     30      3
Oceania          7005     34     35      5      2      0      0      0
Other           34172    309    701     74    161     15      8      1
```

It seems that the data frame contains a few Danish persons without DM or TB diagnoses (mainly with a date of death, though), so we explicitly exclude these persons. Those in this dataset are just a tiny fraction of the group of Danish persons without DM or TB (which constitutes the majority of the Danish population), and whose risk time we shall append later:

```
> dmtb <- subset( dmtb, region !="" | !is.na(dodm) | !is.na(dotb) )
> with( dmtb, ftable( addmargins( table( region,
+                                         Dead=!is.na(dodth),
+                                           DM=!is.na(dodm),
+                                           TB=!is.na(dotb) ),
+                                  margin=1 ),
+                      col.vars=4:2 ) )
          TB    FALSE                        TRUE
          DM    FALSE          TRUE         FALSE          TRUE
          Dead  FALSE   TRUE  FALSE   TRUE  FALSE   TRUE  FALSE   TRUE
region
                    0      0 258154 162034   3479    156    219    148
Africa          31431    409   1262     87   1211     43     71      4
America         47138    275    380     28     26      2      0      0
Asia           122446   1443   5638    522    818     29     74     13
East_Euro      118367   1590   2537    449    172     13     14      4
Europe         243786   2896   4604    517    373     23     30      3
Oceania          7005     34     35      5      2      0      0      0
Other           34172    309    701     74    161     15      8      1
Sum            604345   6956 273311 163716   6242    281    416    173
```

Then we transform dates to date-format, and subsequently transform all date variables in the data frame to `cal.yr` format:

```
> dv <- grep( "do", names(dmtb) )
> names( dmtb )[dv]
[1] "dodm"  "dodth" "dotb"  "doind" "doud"  "dobth"
> for( i in dv ) dmtb[,i] <- as.Date( dmtb[,i], origin="1960-01-01" )
> dmtb$sex <- factor( dmtb$sex, labels=c("M","F") )
> dmtb <- cal.yr( dmtb )
```

We then restrict the data by excluding persons that are dead or emigrated before 1.1.1995 or have no date of birth.

```
> dmtb <- subset( dmtb, pmin( dodth, doud, 1995, na.rm=TRUE ) >= 1995 &
+                       !is.na(dobth) )
> formatC( with( dmtb, addmargins( table( Emigr=!is.na(doud),
+                                         Immigr=!is.na(doind) ) ) ),
+            format="f", big.mark=",", digits=0, pre="common" )
       Immigr
Emigr   FALSE     TRUE       Sum
  FALSE   397,113   377,296   774,409
  TRUE      2,390   251,408   253,798
  Sum     399,503   628,704 1,028,207
```

Not all emigration dates are after immigration dates, so we assume that these are cases of re-immigration, and we decide just to follow these persons from the date of the immigration (`doind`), and ignore the earlier emigration date (`doud`) by setting the latter to `NA`:

```
> dmtb$doud <- with( dmtb, ifelse( doud>doind, doud, NA ) )
> with( dmtb, table( doud>doind, exclude=NULL ) )
  TRUE   <NA>
251358 776849
```

To get an overview of the material, we make histograms of all the date variables, to check whether their ranges and distributions look sensible:

```
> par( mfrow=c(3,2) )
> with( dmtb, hist(dobth,breaks=100, col="gray" ) )
> with( dmtb, hist(dodth,breaks=100, col="gray" ) )
> with( dmtb, hist(doind,breaks=100, col="gray" ) )
> with( dmtb, hist(doud ,breaks=100, col="gray" ) )
> with( dmtb, hist(dodm ,breaks=100, col="gray" ) )
> with( dmtb, hist(dotb ,breaks=100, col="gray" ) )
```

It is clear that the death dates are incomplete beyond 1.1.2010, so we set the end of follow-up to 01.01.2010:

```
> ( end <- cal.yr( as.Date("2010-01-01") ) )
[1] 2010
attr(,"class")
[1] "cal.yr"  "numeric"
```

We can explore the apparent seasonality of the emigration date, by listing those dates that occur more the 400 times in the material:

```
> tt <- table( dmtb$doud )
> names( tt ) <- round( as.numeric(names(tt)), 2 )
> sort( names(tt[tt>400]) )
 [1] "1998.5"  "1999"    "1999.49" "2000.5"  "2001.5"  "2001.58" "2002"    "2002.5"  "2003.49"
[10] "2003.49" "2004"    "2004.49" "2004.5"  "2005"    "2005.49" "2005.5"  "2005.58" "2006"
[19] "2006.42" "2006.49" "2006.5"  "2006.58" "2006.67" "2007"    "2007.42" "2007.49" "2007.49"
[28] "2007.58" "2008"    "2008.02" "2008.08" "2008.41" "2008.49" "2008.5"  "2008.58" "2008.67"
[37] "2009"    "2009"    "2009.09" "2009.42" "2009.49" "2009.5"  "2009.58" "2009.67" "2010"
[46] "2010"    "2010.08" "2010.5"  "2010.58" "2010.67"
```

We also see that there are 589 persons with both a date of DM and of TB, of which 5 have identical values of the two:

```
> with( dmtb, table( dodm==dotb, exclude=NULL ) )
  FALSE    TRUE    <NA>
    584       5 1027618
```

In order to handle the follow-up properly, we define entry and exit dates. Note that we use the `na.rm=TRUE` argument to make sure that we get a valid date for all. Also note that we end follow up at `end` as defined above, and finally adjust the diabetes date to one week prior to TB if it equals the TB date.

```
> dmtb <- transform( dmtb, entry = pmax( dobth, doind, 1995, na.rm=TRUE ),
+                           exit = pmin( dodth, doud ,  end, na.rm=TRUE ),
+                           dodm = pmin( dodm, dodm-(dodm==dotb)/52, na.rm=TRUE) )
> summary( dmtb )
```

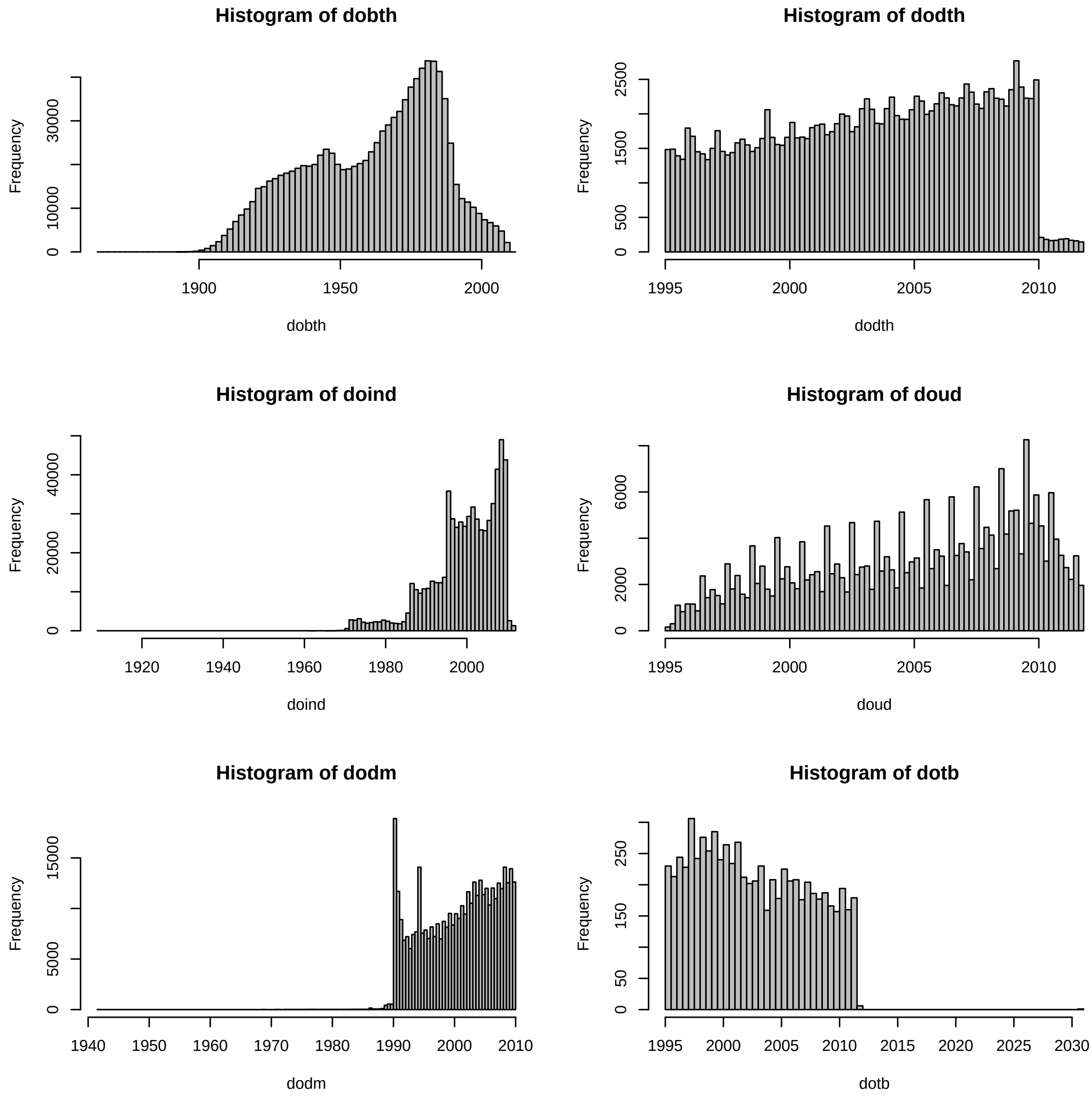

Figure 1.1: *Histograms of all the date variables. The very distinct seasonality of* `doud` *are from the massive over-representation of the dates 1 January and 1 July as seen below.*

```
     dodm            dodth             dotb                region           doind
Min.   :1942     Min.   :1995     Min.   :1995               :397122   Min.   :1910
1st Qu.:1995     1st Qu.:2000     1st Qu.:1999      Europe   :252147   1st Qu.:1996
Median :2001     Median :2004     Median :2002      Asia     :130971   Median :2001
Mean   :2001     Mean   :2003     Mean   :2003      East_Euro:123138   Mean   :2000
3rd Qu.:2006     3rd Qu.:2007     3rd Qu.:2006      America  : 47825   3rd Qu.:2006
Max.   :2010     Max.   :2012     Max.   :2031      Other    : 35414   Max.   :2012
NA's   :617658   NA's   :884149   NA's   :1021096   (Other)  : 41590   NA's   :399503
     doud        sex            dobth          entry           exit
Min.   :1995     M:524770   Min.   :1865   Min.   :1995   Min.   :1995
1st Qu.:2002     F:503437   1st Qu.:1943   1st Qu.:1995   1st Qu.:2007
Median :2006                Median :1966   Median :1995   Median :2010
Mean   :2005                Mean   :1962   Mean   :1999   Mean   :2008
3rd Qu.:2009                3rd Qu.:1981   3rd Qu.:2003   3rd Qu.:2010
```

```
 Max.   :2012               Max.   :2010   Max.   :2012   Max.   :2010  
 NA's   :776849                                                          
```

```
> with( dmtb, table( dodm==dotb, exclude=NULL ) )
```

```
  FALSE    <NA> 
    589 1027618 
```

1.3 Follow-up

We now set up a Lexis object to represent the follow-up; in the first instance just from start till emigration, death or end of follow-up:

```
> Lx <- Lexis( entry = list( date=entry,
+                            age=entry-dobth),
+               exit = list( date=exit ),
+        exit.status = factor( !is.na(dodth), labels=c("Well","Dead") ),
+               data = subset( dmtb, entry<exit ) )
NOTE: entry.status has been set to "Well" for all.
```

```
> summary( Lx )
```

```
Transitions:
     To
From     Well   Dead  Records:  Events: Risk time:  Persons:
  Well 880222 144046   1024268   144046    9208270   1024268
```

We must preserve *both* intermediate events, so we have to cut 6 times:

- At DM where no TB is present
- At TB where no DM is present
- At DM where DM is before TB
- At TB where DM is before TB
- At TB where DM is after TB
- At DM where DM is after TB

To this end we make four data frames with the various combinations of DM and TB dates, and a fifth where no cutting of follow-up is required. The point of this is to separate out the two large parts of the data where no cutting is required (oLx) or only cutting by DM date is required (DM.only). The remaining parts of the cutting are small and require only little computing time:

```
> oLx     <- subset( Lx,  is.na(dodm) &  is.na(dotb) )
> DM.only <- subset( Lx, !is.na(dodm) &  is.na(dotb) )
> TB.only <- subset( Lx,  is.na(dodm) & !is.na(dotb) )
> TB.2nd  <- subset( Lx, !is.na(dodm) & !is.na(dotb) & (dodm < dotb) )
> DM.2nd  <- subset( Lx, !is.na(dodm) & !is.na(dotb) & (dotb < dodm) )
> ( tt <- rbind(dim(oLx    ),
+               dim(DM.only),
+               dim(TB.2nd ),
+               dim(TB.only),
+               dim(DM.2nd )) )
       [,1] [,2]
[1,] 607228   16
[2,] 409933   16
[3,]    360   16
[4,]   6519   16
[5,]    228   16
```

```
> c( nrow(Lx), sum( tt[,1] ) )
```

```
[1] 1024268 1024268
```

We can now cut the follow-up in the different instances and re-assemble afterwards:

```
> system.time(
+  dLx <- cutLexis( DM.only, cut = DM.only$dodm,
+                            pre = "Well",
+                         new.st = "DM",
+                         new.sc = "DMdur" ) )
   user  system elapsed
 110.80    1.73  113.32
```

```
>  tLx <- cutLexis( TB.only, cut = TB.only$dotb,
+                            pre = "Well",
+                         new.st = "TB" )
> dtLx <- cutLexis(  TB.2nd, cut = TB.2nd$dodm,
+                            pre = "Well",
+                         new.st = "DM",
+                         new.sc = "DMdur" )
> dtLx <- cutLexis(    dtLx, cut = dtLx$dotb,
+                            pre = c("Well","DM"),
+                         new.st = "TB(DM)" )
> tdLx <- cutLexis(  DM.2nd, cut = DM.2nd$dotb,
+                            pre = "Well",
+                         new.st = "TB" )
> tdLx <- cutLexis(    tdLx, cut = tdLx$dodm,
+                            pre = c("Well","TB"),
+                         new.st = "DM(TB)" )
```

In assembling the different cut frames we need a function that adds a timescale to a `Lexis` object and just fills it with `NA`s

```
> xsc <-
+ function( x, new.sc )
+ {
+ sc.num <- length( attr(x,"time.scales") )
+ sc.nam <- c( attr( x, "time.scales" ), new.sc )
+ br.new <- c( attr( x, "breaks" ), list( NULL ) )
+ names( br.new ) <- sc.nam
+ xx <- cbind( x[,1:sc.num], as.numeric(NA), x[,-(1:sc.num)] )
+ names( xx )[sc.num+1] <- new.sc
+ attr( xx, "class"       ) <- attr( x, "class" )
+ attr( xx, "time.scales" ) <- sc.nam
+ attr( xx, "breaks"      ) <- br.new
+ xx
+ }
```

There is no need to fidget with the differing factor-levels for `lex.Cst` and `lex.Xst`; this is automatically handled by `rbind`:

```
> xLx <- rbind( xsc(  oLx, "DMdur" ),
+                     dLx,
+               xsc(  tLx, "DMdur" ),
+                    dtLx,
+               xsc( tdLx, "DMdur" ) )
> summary( xLx )
Transitions:
     To
From       Well   Dead     DM    TB TB(DM) DM(TB)  Records:  Events: Risk time:  Persons:
  Well   600816   7186 311225  6143      0      0    925370   324554 6529311.08    925370
  DM          0 136431 273284     0    326      0    410041   136757 2629988.44    410041
  TB          0    258      0  5737      0    228      6223      486   46386.63      6223
  TB(DM)      0    127      0     0    201      0       328      127    1639.06       328
  DM(TB)      0     44      0     0      0    184       228       44     944.41       228
  Sum    600816 144046 584509 11880    527    412   1342190   461968 9208269.61   1024268
```

Once we have cut the follow-up so that we have follow-up through stages, we can show the amount of risk time and the transition rates between the states.

```
> par( mfrow=c(2,1) )
> aclr <- rep("black",8)
> aclr[5] <- "red"
> aclr[3] <- "forestgreen"
> boxes.Lexis( xLx, boxpos=list( x=c(10,90,10,50,50,90),
+                                y=c(65,35,35,90,10,65) ),
+                    hmult=1.5, col.arr=aclr,
+                    scale.Y=1000, scale.R=100 )
> boxes.Lexis( subset(xLx,region!=""),
+               boxpos=list( x=c(10,90,10,50,50,90),
+                            y=c(65,35,35,90,10,65) ),
+               hmult=1.5, col.arr=aclr,
+               scale.Y=1000, scale.R=100 )
```

The rates from the state ”Well” in figure 1.2 are strongly misleading as the persons included here all either contribute to the DM or TB risk time *or* are born outside Denmark. In the display where only the non-Danish born are included, as in figure 1.2, it seems that diabetes is associated with an about 50% increased incidence (from 60 to 90) of tuberculosis.

But that remains to be seen.

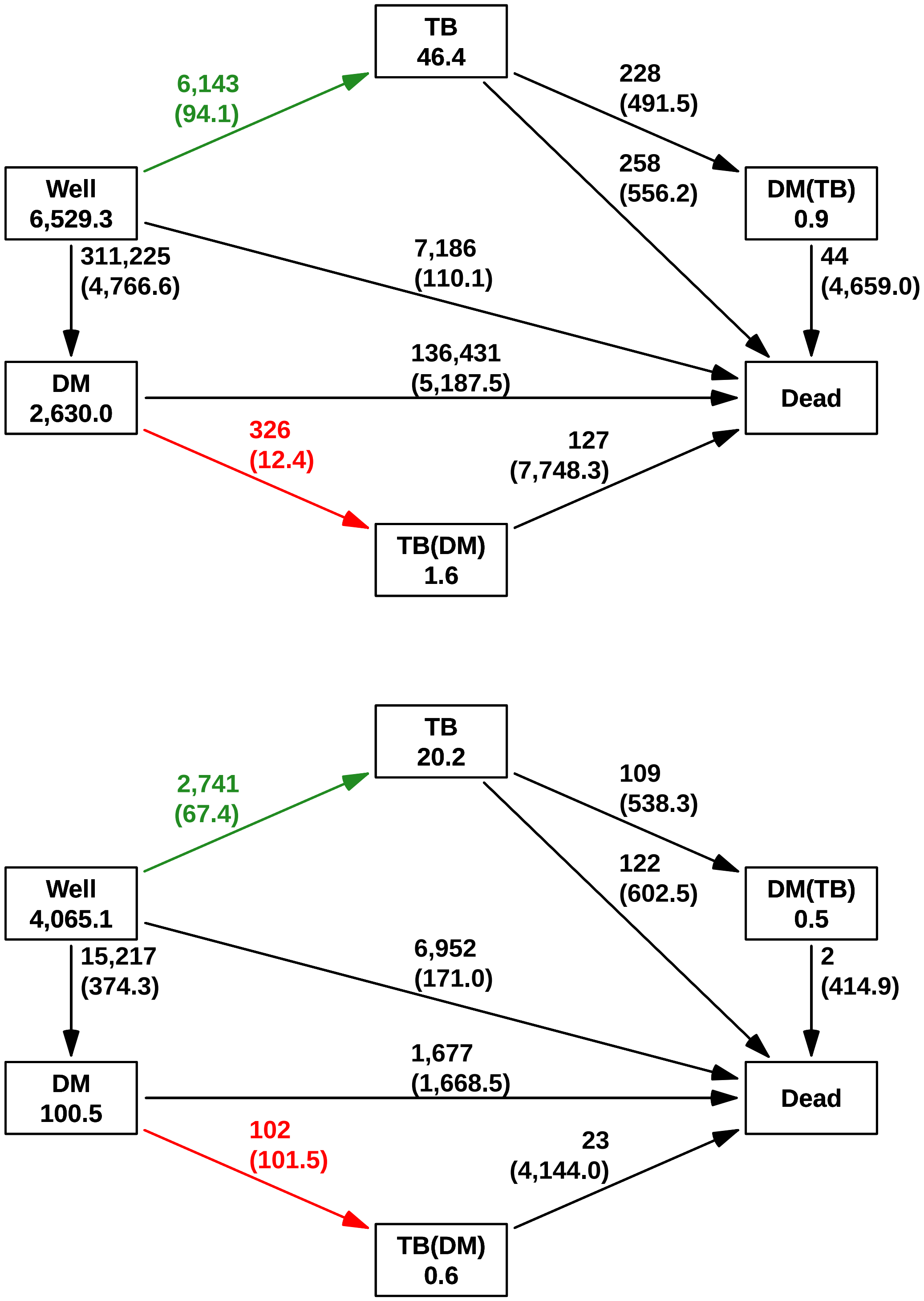

Figure 1.2: *States and transitions between them. Numbers in boxes are person-years in 1000s and numbers on the arrows are number of transitions and transition rates per 100,000 person-years.*

The upper panel shows data from all persons from the database, and thus the rates from the state "Well" are strongly misleading as the persons included here all either contribute to the DM or TB risk time or *are born outside Denmark.*

The lower panel show only data from foreign born persons only, and thus all rates shown are comparable.

The two transition rates that we are interested in comparing are those in red and green.

Chapter 2

Splitting follow-up and adding population data

2.1 Splitting follow-up

We now split the follow-up data by age and calendar time in bands of 1 year in order to classify the risk time among those with diabetes, TB and of foreign birth by sex, age and date of follow-up. We shall subsequently subtract the thus derived risk time from the overall population as obtained from Statistics Denmark, in order to obtain the correct risk time figures for the `Well` state for those born in Denmark..

In practice the time-splitting will produce some 30 intervals per person, so about 30 million intervals, which will not fit into this crap little office computer.

So we split the data for smaller chunks of `xLx` at at time, and aggregate the risk time and TB events into a dataset. This is then merged with and used to update the previous dataset, so we get a sequential updating of events and risk time (as well as a slowly increasing number of rows, as each chunk of the `Lexis` object contains a few combinations of the classifying factors that have not been encountered in previous chunks:

```
> n.chunks <- 100
> lm <- round( seq(0,nrow(xLx),,n.chunks+1) )
> i <- 1
> whr <- (lm[i]+1):(lm[i+1])
> sLx <- splitLexis( xLx[whr,], 0:100, time.scale="age" )
> sLx <- splitLexis( sLx,  1995:2012, time.scale="date" )
> Agg <- with( sLx, aggregate( cbind( Y = lex.dur,
+                                   D.tb = ( lex.Xst %in% c("TB","TB(DM)") &
+                                            lex.Xst != lex.Cst )*1,
+                                   D.dm = ( lex.Xst %in% c("DM","DM(TB)") &
+                                            lex.Xst != lex.Cst )*1,
+                                   D.dd = ( lex.Xst == "Dead" )*1 ),
+                               list( A = floor(age),
+                                     P = floor(date),
+                                     U = floor(date)-floor(age)-floor(dobth),
+                                   sex = sex,
+                                region = region,
+                                 state = lex.Cst ),
+                               FUN = sum ) )
> c( nrow(sLx ), nrow( Agg ) )
[1] 219594  22801
> for( i in 2:n.chunks )
+ {
+ whr <- (lm[i]+1):(lm[i+1])
+ sLx <- splitLexis( xLx[whr,], 0:100, time.scale="age" )
+ sLx <- splitLexis( sLx,  1995:2012, time.scale="date" )
+ agg <- with( sLx, aggregate( cbind( y = lex.dur,
+                                  d.tb = ( lex.Xst %in% c("TB","TB(DM)") &
+                                           lex.Xst != lex.Cst )*1,
+                                  d.dm = ( lex.Xst %in% c("DM","DM(TB)") &
+                                           lex.Xst != lex.Cst )*1,
+                                  d.dd = ( lex.Xst == "Dead" )*1 ),
+                              list( A = floor(age),
+                                    P = floor(date),
+                                    U = floor(date)-floor(age)-floor(dobth),
+                                  sex = sex,
+                               region = region,
+                                state = lex.Cst ),
+                              FUN = sum ) )
+ Agg <- merge( Agg, agg, by=names( Agg )[1:6], all=TRUE )
+ Agg <- transform( Agg, Y = pmax(Y   ,0,na.rm=TRUE) + pmax(y   ,0,na.rm=TRUE),
+                     D.tb = pmax(D.tb,0,na.rm=TRUE) + pmax(d.tb,0,na.rm=TRUE),
+                     D.dm = pmax(D.dm,0,na.rm=TRUE) + pmax(d.dm,0,na.rm=TRUE),
+                     D.dd = pmax(D.dd,0,na.rm=TRUE) + pmax(d.dd,0,na.rm=TRUE) )[,
+                   c("A","P","U","sex","region","state","Y","D.tb","D.dm","D.dd")]
+ cat( "Merged in chunk", i, "now", nrow(Agg), "rows, at",
+      format(Sys.time(),format="%Y-%m-%d %H:%M:%S"), "\n" )
+ }
```

```
Merged in chunk 2 now 27311 rows, at 2013-06-27 22:58:47
Merged in chunk 3 now 28953 rows, at 2013-06-27 22:59:12
Merged in chunk 4 now 29963 rows, at 2013-06-27 22:59:34
Merged in chunk 5 now 31049 rows, at 2013-06-27 22:59:58
Merged in chunk 6 now 31754 rows, at 2013-06-27 23:00:20
Merged in chunk 7 now 32356 rows, at 2013-06-27 23:00:44
Merged in chunk 8 now 32752 rows, at 2013-06-27 23:01:08
Merged in chunk 9 now 33209 rows, at 2013-06-27 23:01:32
Merged in chunk 10 now 33608 rows, at 2013-06-27 23:01:56
Merged in chunk 11 now 33945 rows, at 2013-06-27 23:02:20
Merged in chunk 12 now 34200 rows, at 2013-06-27 23:02:42
Merged in chunk 13 now 34401 rows, at 2013-06-27 23:03:05
Merged in chunk 14 now 34659 rows, at 2013-06-27 23:03:27
Merged in chunk 15 now 34821 rows, at 2013-06-27 23:03:53
Merged in chunk 16 now 34971 rows, at 2013-06-27 23:04:19
Merged in chunk 17 now 35054 rows, at 2013-06-27 23:04:41
Merged in chunk 18 now 35265 rows, at 2013-06-27 23:05:04
Merged in chunk 19 now 35367 rows, at 2013-06-27 23:05:25
Merged in chunk 20 now 35472 rows, at 2013-06-27 23:05:47
Merged in chunk 21 now 35582 rows, at 2013-06-27 23:06:08
Merged in chunk 22 now 35687 rows, at 2013-06-27 23:06:30
Merged in chunk 23 now 35790 rows, at 2013-06-27 23:06:53
Merged in chunk 24 now 35975 rows, at 2013-06-27 23:07:17
Merged in chunk 25 now 36060 rows, at 2013-06-27 23:07:37
Merged in chunk 26 now 36159 rows, at 2013-06-27 23:07:59
Merged in chunk 27 now 36211 rows, at 2013-06-27 23:08:21
Merged in chunk 28 now 36328 rows, at 2013-06-27 23:08:43
Merged in chunk 29 now 36345 rows, at 2013-06-27 23:09:04
Merged in chunk 30 now 36422 rows, at 2013-06-27 23:09:29
Merged in chunk 31 now 36544 rows, at 2013-06-27 23:09:52
Merged in chunk 32 now 36572 rows, at 2013-06-27 23:10:16
Merged in chunk 33 now 36648 rows, at 2013-06-27 23:10:39
Merged in chunk 34 now 36696 rows, at 2013-06-27 23:11:01
Merged in chunk 35 now 36806 rows, at 2013-06-27 23:11:25
Merged in chunk 36 now 36831 rows, at 2013-06-27 23:11:47
Merged in chunk 37 now 36874 rows, at 2013-06-27 23:12:09
Merged in chunk 38 now 36959 rows, at 2013-06-27 23:12:32
Merged in chunk 39 now 37020 rows, at 2013-06-27 23:12:53
Merged in chunk 40 now 37052 rows, at 2013-06-27 23:13:14
Merged in chunk 41 now 37094 rows, at 2013-06-27 23:13:36
Merged in chunk 42 now 37158 rows, at 2013-06-27 23:13:59
Merged in chunk 43 now 37240 rows, at 2013-06-27 23:14:20
Merged in chunk 44 now 37298 rows, at 2013-06-27 23:14:42
Merged in chunk 45 now 37343 rows, at 2013-06-27 23:15:04
Merged in chunk 46 now 52477 rows, at 2013-06-27 23:15:30
Merged in chunk 47 now 56882 rows, at 2013-06-27 23:15:53
Merged in chunk 48 now 58700 rows, at 2013-06-27 23:16:17
Merged in chunk 49 now 59541 rows, at 2013-06-27 23:16:42
Merged in chunk 50 now 60340 rows, at 2013-06-27 23:17:07
Merged in chunk 51 now 60988 rows, at 2013-06-27 23:17:33
Merged in chunk 52 now 61569 rows, at 2013-06-27 23:17:56
Merged in chunk 53 now 62141 rows, at 2013-06-27 23:18:20
Merged in chunk 54 now 62698 rows, at 2013-06-27 23:18:46
Merged in chunk 55 now 63110 rows, at 2013-06-27 23:19:11
Merged in chunk 56 now 63638 rows, at 2013-06-27 23:19:35
Merged in chunk 57 now 64100 rows, at 2013-06-27 23:19:59
Merged in chunk 58 now 64486 rows, at 2013-06-27 23:20:24
Merged in chunk 59 now 64739 rows, at 2013-06-27 23:20:48
Merged in chunk 60 now 65058 rows, at 2013-06-27 23:21:13
Merged in chunk 61 now 65334 rows, at 2013-06-27 23:21:39
Merged in chunk 62 now 65631 rows, at 2013-06-27 23:22:05
Merged in chunk 63 now 65962 rows, at 2013-06-27 23:22:29
Merged in chunk 64 now 66166 rows, at 2013-06-27 23:22:54
Merged in chunk 65 now 66407 rows, at 2013-06-27 23:23:18
Merged in chunk 66 now 66650 rows, at 2013-06-27 23:23:42
Merged in chunk 67 now 66977 rows, at 2013-06-27 23:24:08
Merged in chunk 68 now 67108 rows, at 2013-06-27 23:24:32
Merged in chunk 69 now 67310 rows, at 2013-06-27 23:24:58
Merged in chunk 70 now 67544 rows, at 2013-06-27 23:25:23
Merged in chunk 71 now 67817 rows, at 2013-06-27 23:25:48
Merged in chunk 72 now 68187 rows, at 2013-06-27 23:26:13
Merged in chunk 73 now 68348 rows, at 2013-06-27 23:26:37
Merged in chunk 74 now 68511 rows, at 2013-06-27 23:27:01
Merged in chunk 75 now 68694 rows, at 2013-06-27 23:27:25
Merged in chunk 76 now 68832 rows, at 2013-06-27 23:27:51
Merged in chunk 77 now 68908 rows, at 2013-06-27 23:28:14
Merged in chunk 78 now 69023 rows, at 2013-06-27 23:28:38
Merged in chunk 79 now 69211 rows, at 2013-06-27 23:29:03
Merged in chunk 80 now 69411 rows, at 2013-06-27 23:29:27
Merged in chunk 81 now 69579 rows, at 2013-06-27 23:29:51
Merged in chunk 82 now 69701 rows, at 2013-06-27 23:30:16
Merged in chunk 83 now 69774 rows, at 2013-06-27 23:30:40
Merged in chunk 84 now 69865 rows, at 2013-06-27 23:31:06
Merged in chunk 85 now 70025 rows, at 2013-06-27 23:31:30
Merged in chunk 86 now 70132 rows, at 2013-06-27 23:31:54
Merged in chunk 87 now 70268 rows, at 2013-06-27 23:32:18
Merged in chunk 88 now 70463 rows, at 2013-06-27 23:32:43
Merged in chunk 89 now 70560 rows, at 2013-06-27 23:33:07
Merged in chunk 90 now 70670 rows, at 2013-06-27 23:33:32
Merged in chunk 91 now 70819 rows, at 2013-06-27 23:33:56
Merged in chunk 92 now 70868 rows, at 2013-06-27 23:34:20
Merged in chunk 93 now 70966 rows, at 2013-06-27 23:34:44
Merged in chunk 94 now 71196 rows, at 2013-06-27 23:35:08
Merged in chunk 95 now 71288 rows, at 2013-06-27 23:35:33
Merged in chunk 96 now 71418 rows, at 2013-06-27 23:36:00
Merged in chunk 97 now 71487 rows, at 2013-06-27 23:36:23
Merged in chunk 98 now 71588 rows, at 2013-06-27 23:36:48
Merged in chunk 99 now 75227 rows, at 2013-06-27 23:37:12
Merged in chunk 100 now 92219 rows, at 2013-06-27 23:37:37
> summary( Agg )
       A                P              U          sex             region    
 Min.   :  0.00   Min.   :1995   Min.   :0.0000   M:45432            :19926  
 1st Qu.: 27.00   1st Qu.:1999   1st Qu.:0.0000   F:46787   Asia     :13862  
 Median : 46.00   Median :2003   Median :0.0000             Europe   :12728  
 Mean   : 46.63   Mean   :2002   Mean   :0.4966             Africa   :12114  
 3rd Qu.: 66.00   3rd Qu.:2006   3rd Qu.:1.0000             East_Euro:11464  
 Max.   :144.00   Max.   :2009   Max.   :1.0000             Other    : 9459  
                                                            (Other)  :12666  
```

```
    state             Y                   D.tb                D.dm
 Well  :43481   Min.   :   0.0007   Min.    :0.00000   Min.   :  0.000
 Dead  :    0   1st Qu.:   1.1756   1st Qu.:0.00000   1st Qu.:  0.000
 DM    :28187   Median :   6.3943   Median :0.00000   Median :  0.000
 TB    :16373   Mean   :  99.8522   Mean    :0.07015   Mean   :  3.377
 TB(DM): 2536   3rd Qu.:  64.5565   3rd Qu.:0.00000   3rd Qu.:  0.000
 DM(TB): 1642   Max.   :2523.1027   Max.    :7.00000   Max.   :262.000

      D.dd
 Min.   :  0.000
 1st Qu.:  0.000
 Median :  0.000
 Mean   :  1.562
 3rd Qu.:  0.000
 Max.   :134.000
```

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

2.2 Splitting follow-up by duration

We will also be splitting the follow-up among those with diabetes by diabetes duration, however, only for diabetes patients diagnosed after 1.1.1995:

```
> dLx <- subset( xLx, lex.Cst=="DM" & dodm>1995 )
> with( dLx, table( lex.Xst ) )
lex.Xst
  Well   Dead     DM     TB TB(DM) DM(TB)
     0  76790 234432      0    223      0
> nrow( dLx )
[1] 311445
```

The code to complete this task is almost the same as before, except that we have included diabetes duration in fairly small intervals, and by that token made a shortcut in the splitting, as we only split by diabetes duration, and just classify follow-up according to where it belongs:

```
> n.chunks <- 50
> lm <- round( seq(0,nrow(dLx),,n.chunks+1) )
> i <- 1
> whr <- (lm[i]+1):(lm[i+1])
> sLx <- splitLexis( dLx[whr,], breaks=seq(0,20,0.2), time.scale="DMdur" )
> Dgg <- with( sLx,
+              aggregate( cbind( Y = lex.dur,
+                             D.tb = ( lex.Xst %in% c("TB","TB(DM)") &
+                                      lex.Xst != lex.Cst )*1,
+                             D.dm = ( lex.Xst %in% c("DM","DM(TB)") &
+                                      lex.Xst != lex.Cst )*1,
+                             D.dd = ( lex.Xst == "Dead" )*1 ),
+                         list( A = floor(age+0.1),
+                               P = floor(date+0.1),
+                               U = floor(date+0.1)-floor(age+0.1)-floor(dobth),
+                             dur = timeBand( sLx, "DMdur", "left" ),
+                             sex = sex,
+                          region = region,
+                           state = lex.Cst ),
+                         FUN = sum ) )
> c( nrow(sLx ), nrow( Dgg ) )
[1] 175309 103016
```

```
> for( i in 2:n.chunks )
+ {
+ whr <- (lm[i]+1):(lm[i+1])
+ sLx <- splitLexis( dLx[whr,], breaks=seq(0,20,0.2), time.scale="DMdur" )
+ dgg <- with( sLx,
+              aggregate( cbind( y = lex.dur,
+                             d.tb = ( lex.Xst %in% c("TB","TB(DM)") &
+                                      lex.Xst != lex.Cst )*1,
+                             d.dm = ( lex.Xst %in% c("DM","DM(TB)") &
+                                      lex.Xst != lex.Cst )*1,
+                             d.dd = ( lex.Xst == "Dead" )*1 ),
+                         list( A = floor(age+0.1),
+                               P = floor(date+0.1),
+                               U = floor(date+0.1)-floor(age+0.1)-floor(dobth),
+                             dur = timeBand( sLx, "DMdur", "left" ),
+                             sex = sex,
+                          region = region,
+                           state = lex.Cst ),
+                         FUN = sum ) )
+ Dgg <- merge( Dgg, dgg, by=names( Dgg )[1:7], all=TRUE )
+ Dgg <- transform( Dgg, Y = pmax(Y   ,0,na.rm=TRUE) + pmax(y   ,0,na.rm=TRUE),
+                     D.tb = pmax(D.tb,0,na.rm=TRUE) + pmax(d.tb,0,na.rm=TRUE),
+                     D.dm = pmax(D.dm,0,na.rm=TRUE) + pmax(d.dm,0,na.rm=TRUE),
+                     D.dd = pmax(D.dd,0,na.rm=TRUE) + pmax(d.dd,0,na.rm=TRUE) )[,
+                   c("A","P","U","dur","sex","region","state","Y","D.tb","D.dm","D.dd")]
+ cat( "Merged in chunk", i, "now", nrow(Dgg), "rows, at",
+      format(Sys.time(),format="%Y-%m-%d %H:%M:%S"), "\n" )
+ }
```

```
Merged in chunk 2 now 164657 rows, at 2013-06-27 23:38:19
Merged in chunk 3 now 193220 rows, at 2013-06-27 23:38:36
Merged in chunk 4 now 214047 rows, at 2013-06-27 23:38:54
Merged in chunk 5 now 230563 rows, at 2013-06-27 23:39:12
Merged in chunk 6 now 244506 rows, at 2013-06-27 23:39:30
Merged in chunk 7 now 257264 rows, at 2013-06-27 23:39:48
Merged in chunk 8 now 271620 rows, at 2013-06-27 23:40:07
Merged in chunk 9 now 282168 rows, at 2013-06-27 23:40:26
Merged in chunk 10 now 292284 rows, at 2013-06-27 23:40:46
Merged in chunk 11 now 302111 rows, at 2013-06-27 23:41:05
Merged in chunk 12 now 310283 rows, at 2013-06-27 23:41:24
Merged in chunk 13 now 318213 rows, at 2013-06-27 23:41:44
Merged in chunk 14 now 324765 rows, at 2013-06-27 23:42:03
Merged in chunk 15 now 330960 rows, at 2013-06-27 23:42:23
Merged in chunk 16 now 339207 rows, at 2013-06-27 23:42:43
Merged in chunk 17 now 349367 rows, at 2013-06-27 23:43:03
Merged in chunk 18 now 355335 rows, at 2013-06-27 23:43:24
Merged in chunk 19 now 362387 rows, at 2013-06-27 23:43:44
Merged in chunk 20 now 369445 rows, at 2013-06-27 23:44:05
Merged in chunk 21 now 374207 rows, at 2013-06-27 23:44:26
Merged in chunk 22 now 379552 rows, at 2013-06-27 23:44:46
Merged in chunk 23 now 384341 rows, at 2013-06-27 23:45:07
Merged in chunk 24 now 391193 rows, at 2013-06-27 23:45:28
Merged in chunk 25 now 399000 rows, at 2013-06-27 23:45:48
Merged in chunk 26 now 404340 rows, at 2013-06-27 23:46:10
Merged in chunk 27 now 409944 rows, at 2013-06-27 23:46:32
Merged in chunk 28 now 414250 rows, at 2013-06-27 23:46:53
Merged in chunk 29 now 418858 rows, at 2013-06-27 23:47:14
Merged in chunk 30 now 423301 rows, at 2013-06-27 23:47:36
Merged in chunk 31 now 427329 rows, at 2013-06-27 23:47:57
Merged in chunk 32 now 432821 rows, at 2013-06-27 23:48:20
Merged in chunk 33 now 438654 rows, at 2013-06-27 23:48:42
Merged in chunk 34 now 443570 rows, at 2013-06-27 23:49:03
Merged in chunk 35 now 447412 rows, at 2013-06-27 23:49:25
Merged in chunk 36 now 450749 rows, at 2013-06-27 23:49:48
Merged in chunk 37 now 454790 rows, at 2013-06-27 23:50:10
Merged in chunk 38 now 458620 rows, at 2013-06-27 23:50:32
Merged in chunk 39 now 462715 rows, at 2013-06-27 23:50:53
Merged in chunk 40 now 466638 rows, at 2013-06-27 23:51:16
Merged in chunk 41 now 470551 rows, at 2013-06-27 23:51:39
Merged in chunk 42 now 474442 rows, at 2013-06-27 23:52:02
Merged in chunk 43 now 477591 rows, at 2013-06-27 23:52:24
Merged in chunk 44 now 480714 rows, at 2013-06-27 23:52:47
Merged in chunk 45 now 484346 rows, at 2013-06-27 23:53:10
Merged in chunk 46 now 487799 rows, at 2013-06-27 23:53:33
Merged in chunk 47 now 490901 rows, at 2013-06-27 23:53:56
Merged in chunk 48 now 494034 rows, at 2013-06-27 23:54:19
Merged in chunk 49 now 496913 rows, at 2013-06-27 23:54:42
Merged in chunk 50 now 500576 rows, at 2013-06-27 23:55:04
```

```
> str( Dgg )
```

```
'data.frame':        500576 obs. of  11 variables:
 $ A     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ P     : num  1995 1995 1995 1995 1996 ...
 $ U     : num  0 0 0 0 0 0 0 0 1 1 ...
 $ dur   : num  0 0.2 0.4 0.6 0 0.2 0.4 0.6 0 0.2 ...
 $ sex   : Factor w/ 2 levels "M","F": 1 1 1 1 2 2 2 2 2 2 ...
 $ region: Factor w/ 8 levels "","Africa","America",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ state : Factor w/ 6 levels "Well","Dead",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ Y     : num  0.2 0.2 0.2 0.2 0.4 0.4 0.4 0.2 0.6 0.4 ...
 $ D.tb  : num  0 0 0 0 0 0 0 0 0 0 ...
 $ D.dm  : num  0 0 0 0 0 0 0 0 0 0 ...
 $ D.dd  : num  0 0 0 0 0 0 0 0 0 0 ...
```

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

2.3 Acquiring the population risk time

So far we have only attended to persons who are either non-Danish or have a diagnosis of DM or TB. So in the “Well” state we are missing the follow-up time from Danish persons without DM or TB. But we actually have access to all other follow-up time in the object `Agg`, so if we take this risk time and subtract from the total risk time in the population, we get the the risk time among Danish in the state “Well”.

```
R version 3.0.1 (2013-05-16)
Platform: i386-w64-mingw32/i386 (32-bit)

attached base packages:
[1] utils     datasets  graphics  grDevices stats     methods   base

other attached packages:
[1] Epi_1.1.51     foreign_0.8-53

loaded via a namespace (and not attached):
[1] tools_3.0.1
```

The data frame `Agg` contains all the risk time among the persons on whom we have follow-up in the various states.

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

```
'data.frame':        92219 obs. of  10 variables:
 $ A     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ P     : num  1995 1995 1995 1995 1995 ...
 $ U     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ sex   : Factor w/ 2 levels "M","F": 1 1 1 1 1 1 1 1 1 1 ...
 $ region: Factor w/ 8 levels "","Africa","America",..: 1 1 1 2 3 4 5 6 7 8 ...
 $ state : Factor w/ 6 levels "Well","Dead",..: 1 3 4 1 1 1 1 1 1 1 ...
 $ Y     : num  71.919 0.806 0.144 4.534 20.491 ...
 $ D.tb  : num  1 0 0 0 0 0 0 0 0 0 ...
 $ D.dm  : num  1 0 0 0 0 0 0 0 0 0 ...
 $ D.dd  : num  0 0 0 0 0 0 0 0 0 0 ...
```

```
> round(
+ ftable( xtabs( Y/1000 ~ region + state,
+               data = Agg ),
+         row.vars=c(1) ) , 1 )
```

```
          state   Well   Dead     DM     TB TB(DM) DM(TB)
region
                2464.2    0.0 2529.5   26.1    1.1    0.5
Africa           306.6    0.0    8.3   10.1    0.1    0.2
America          230.6    0.0    2.0    0.2    0.0    0.0
Asia            1115.9    0.0   40.0    5.2    0.3    0.1
East_Euro        703.8    0.0   17.4    1.2    0.0    0.0
Europe          1539.2    0.0   28.5    2.6    0.1    0.1
Oceania           31.2    0.0    0.2    0.0    0.0    0.0
Other            137.7    0.0    4.1    0.9    0.0    0.0
```

```
> ftable( xtabs( cbind( D.tb, D.dm, D.dd ) ~ region + state,
+               data = Agg ),
+         row.vars=c(3,1) )
```

```
               state   Well   Dead     DM     TB TB(DM) DM(TB)
     region
D.tb                   3402      0    224      0      0      0
     Africa            1242      0     18      0      0      0
     America             24      0      0      0      0      0
     Asia               782      0     51      0      0      0
     East_Euro          161      0     10      0      0      0
     Europe             373      0     16      0      0      0
     Oceania              1      0      0      0      0      0
     Other              158      0      7      0      0      0
D.dm                 296008      0      0    119      0      0
     Africa            1250      0      0     55      0      0
     America            361      0      0      0      0      0
     Asia              5523      0      0     30      0      0
     East_Euro         2823      0      0      7      0      0
     Europe            4520      0      0     16      0      0
     Oceania             36      0      0      0      0      0
     Other              704      0      0      1      0      0
D.dd                    234      0 134754    136    104     42
     Africa             410      0     87     42      3      1
     America            275      0     28      2      0      0
     Asia              1441      0    520     29     13      0
     East_Euro         1590      0    448     13      3      1
     Europe            2893      0    515     21      3      0
     Oceania             34      0      5      0      0      0
     Other              309      0     74     15      1      0
```

The follow-up time for persons in region `""` and state `"Well"` is wrong, because the dataset should only include persons who either are born outside DK or have either a DM or TB event recorded. Risk time in all other states is correct, and *all* transitions to DM and TB are correct.

But the number of TB and DM events from this state is correct, as we included everyone with any of these events.

There is of course a lot of deaths missing, so for mortality analyses, further expansion of data is required. However, one problem is that we do not have deaths available in Lexis triangles, and anyway mortality analyses are outside the scope of this study.

Thus this risk time computed in `Agg` should be replaced by the total population risk time *minus* the risk time accumulated by those born outside of Denmark *or* by persons with a previous diagnosis of DM or TB. This can be obtained by subtracting from the total population risk time all risk time among persons born outside Denmark *plus* risk time among persons born in Denmark *after* either diagnosis of DM or TB.

But this risk time is readily available in the dataframe of aggregated follow-up, we just sum over the states subsequent to the state “Well”, or among persons not in Denmark (`region==""`)

```
> system.time(
+ Cgg <- with( subset( Agg, A<100 & P>1994 & P<2010 &
+                           !(region=="" & state=="Well") ),
+              aggregate( cbind( X = Y ),
+                         list( A = A,
+                               P = P,
+                           upper = U,
+                             sex = sex ),
+                         FUN = sum ) ) )
   user  system elapsed
   0.99    0.03    1.05
```

```
> str( Cgg )
'data.frame':        6000 obs. of  5 variables:
 $ A    : num  0 1 2 3 4 5 6 7 8 9 ...
 $ P    : num  1995 1995 1995 1995 1995 ...
 $ upper: num  0 0 0 0 0 0 0 0 0 0 ...
 $ sex  : Factor w/ 2 levels "M","F": 1 1 1 1 1 1 1 1 1 1 ...
 $ X    : num  88.3 250.8 282.5 388.1 448.2 ...
```

```
> summary( Cgg )
       A               P             upper      sex            X
 Min.   : 0.00   Min.   :1995   Min.   :0.0   M:3000   Min.   :   0.5455
 1st Qu.:24.75   1st Qu.:1998   1st Qu.:0.0   F:3000   1st Qu.: 628.3542
 Median :49.50   Median :2002   Median :0.5            Median :1010.2361
 Mean   :49.50   Mean   :2002   Mean   :0.5            Mean   :1123.8882
 3rd Qu.:74.25   3rd Qu.:2006   3rd Qu.:1.0            3rd Qu.:1602.2924
 Max.   :99.00   Max.   :2009   Max.   :1.0            Max.   :3135.0233
```

`Cgg` now has the number of person-years lived by persons who are either non-Danish or who have a diagnosis of TB and/or DM, classified by sex and Lexis triangles (age, period and cohort).

Then we get the population data from Denmark in Lexis triangles:

```
> data( Y.dk )
> Y.dk$sex <- factor( Y.dk$sex, labels=c("M","F") )
> Y.dk <- subset( Y.dk,
+                 A<100 & P>1994 & P<2010,
+                 select=c("sex","A","P","upper","Y") )
```

In `Y.dk` we now have the total person-years in the population (up to age 100), and can now subtract the person-years from the study in order to get the follow-up among the non-foreign, non-TB, non-DM persons:

```
> Y.rev <- merge( Cgg, Y.dk, all.y=TRUE )
> summary( Y.rev )
       A               P             upper      sex            X
 Min.   : 0.00   Min.   :1995   Min.   :0.0   M:3000   Min.   :   0.5455
 1st Qu.:24.75   1st Qu.:1998   1st Qu.:0.0   F:3000   1st Qu.: 628.3542
 Median :49.50   Median :2002   Median :0.5            Median :1010.2361
 Mean   :49.50   Mean   :2002   Mean   :0.5            Mean   :1123.8882
 3rd Qu.:74.25   3rd Qu.:2006   3rd Qu.:1.0            3rd Qu.:1602.2924
 Max.   :99.00   Max.   :2009   Max.   :1.0            Max.   :3135.0233
       Y
 Min.   :   48.5
 1st Qu.: 9145.0
 Median :15918.9
 Mean   :13435.6
 3rd Qu.:18404.8
 Max.   :23096.3
```

```
> Y.rev <- transform( Y.rev, Y.pop = Y-pmax(X,0,na.rm=TRUE),
+                             state = "Well",
+                            region = "",
+                                 U = upper )[,c("A","P","U","sex","state","region","Y.pop")]
> str( Y.rev )
```

```
'data.frame':        6000 obs. of  7 variables:
 $ A     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ P     : num  1995 1995 1995 1995 1996 ...
 $ U     : num  0 0 1 1 0 0 1 1 0 0 ...
 $ sex   : Factor w/ 2 levels "M","F": 2 1 2 1 2 1 2 1 2 1 ...
 $ state : Factor w/ 1 level "Well": 1 1 1 1 1 1 1 1 1 1 ...
 $ region: Factor w/ 1 level "": 1 1 1 1 1 1 1 1 1 1 ...
 $ Y.pop : num  16937 17939 16958 17724 16384 ...
```

Thus `Y.rev` now contains the correct person-years in the "Well" state among persons born in DK (`region=""`), classified by sex, age, date of follow-up and date of birth.

2.3.1 Creating follow-up for all persons

The trick is now to merge the new population data in the data frame with the aggregate person-years; first we do this for the dataset with aggregate figures for the entire follow-up:

```
> Afu <- merge( subset( Agg, A<100 & P>1994 & P<2010 ), Y.rev, all=TRUE )
> Afu <- transform( Afu, Y = pmax( Y,Y.pop,na.rm=TRUE),
+                     D.tb = pmax(D.tb,  0,na.rm=TRUE),
+                     D.dm = pmax(D.dm,  0,na.rm=TRUE),
+                     D.dd = pmax(D.dd,  0,na.rm=TRUE) )[,
+                   c("sex","A","P","U","state","region","Y","D.tb","D.dm","D.dd")]
> str( Afu )
```

```
'data.frame':        91922 obs. of  10 variables:
 $ sex   : Factor w/ 2 levels "M","F": 1 1 1 1 1 1 1 1 1 1 ...
 $ A     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ P     : num  1995 1995 1995 1995 1995 ...
 $ U     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ state : Factor w/ 6 levels "Well","Dead",..: 1 3 4 1 1 1 1 1 1 1 ...
 $ region: Factor w/ 8 levels "","Africa","America",..: 1 1 1 2 3 4 5 6 7 8 ...
 $ Y     : num  1.79e+04 8.06e-01 1.44e-01 4.53 2.05e+01 ...
 $ D.tb  : num  1 0 0 0 0 0 0 0 0 0 ...
 $ D.dm  : num  1 0 0 0 0 0 0 0 0 0 ...
 $ D.dd  : num  0 0 0 0 0 0 0 0 0 0 ...
```

The data frame `Afu` now contains the correct number of person-years and transitions to TB and DM , but not to death:

```
> round( addmargins( xtabs( Y ~ region + state, data=Afu )/1000 ), 1 )
```

```
           state
region         Well    Dead      DM      TB  TB(DM)  DM(TB)     Sum
            73870.1     0.0  2528.9    26.1     1.1     0.5 76426.6
  Africa      306.6     0.0     8.3    10.1     0.1     0.2   325.4
  America     230.6     0.0     2.0     0.2     0.0     0.0   232.9
  Asia       1115.9     0.0    40.0     5.2     0.3     0.1  1161.6
  East_Euro   703.8     0.0    17.4     1.2     0.0     0.0   722.5
  Europe     1539.2     0.0    28.5     2.6     0.1     0.1  1570.4
  Oceania      31.2     0.0     0.2     0.0     0.0     0.0    31.3
  Other       137.7     0.0     4.1     0.9     0.0     0.0   142.8
  Sum       77935.2     0.0  2629.4    46.3     1.6     0.9 80613.5
```

```
> ftable( addmargins( xtabs( cbind(D.tb,D.dm) ~ region + state,
+                             data=Afu ),
+                      margin = 1:2 ),
+          row.vars=c(3,1) )
               state   Well   Dead     DM     TB TB(DM) DM(TB)    Sum
     region
D.tb                   3401      0    224      0      0      0   3625
     Africa            1242      0     18      0      0      0   1260
     America             24      0      0      0      0      0     24
     Asia               782      0     51      0      0      0    833
     East_Euro          161      0     10      0      0      0    171
     Europe             373      0     16      0      0      0    389
     Oceania              1      0      0      0      0      0      1
     Other              158      0      7      0      0      0    165
     Sum               6142      0    326      0      0      0   6468
D.dm                 295957      0      0    119      0      0 296076
     Africa            1250      0      0     55      0      0   1305
     America            361      0      0      0      0      0    361
     Asia              5523      0      0     30      0      0   5553
     East_Euro         2823      0      0      7      0      0   2830
     Europe            4520      0      0     16      0      0   4536
     Oceania             36      0      0      0      0      0     36
     Other              704      0      0      1      0      0    705
     Sum             311174      0      0    228      0      0 311402
> save( Afu, file="../data/Afu.Rda" )
```

2.3.2 Merging with duration-classified data

We can now do the same with the follow-up restricted to diabetes patients diagnosed after 1995, that is where all patients diagnosed with DM before 1995 are excluded from follow-up. Thus the dataset Dgg contains no follow-up for the non-diabetic part if the population, neither in terms of person-years nor events of any type:

```
> load( file="../data/Dgg.Rda" )
> with( Dgg, table( state, D.tb ) )
        D.tb
state         0      1
  Well        0      0
  Dead        0      0
  DM     500353    223
  TB          0      0
  TB(DM)      0      0
  DM(TB)      0      0
> with( Dgg, table( state, D.dm ) )
        D.dm
state         0
  Well        0
  Dead        0
  DM     500576
  TB          0
  TB(DM)      0
  DM(TB)      0
> with( Dgg, table( state, D.dd ) )
        D.dd
state         0      1      2      3      4      5      6      7      8      9     10
  Well        0      0      0      0      0      0      0      0      0      0      0
  Dead        0      0      0      0      0      0      0      0      0      0      0
  DM     449249  34155  11555   3858   1167    384    137     44     17      4      5
  TB          0      0      0      0      0      0      0      0      0      0      0
  TB(DM)      0      0      0      0      0      0      0      0      0      0      0
  DM(TB)      0      0      0      0      0      0      0      0      0      0      0
        D.dd
state        11
  Well        0
  Dead        0
  DM          1
  TB          0
  TB(DM)      0
  DM(TB)      0
```

Therefore, we must append the entire follow-up (both person-years and events) as constructed above:

```
> Dfu <- rbind( subset( Dgg, A<100 ),
+        cbind( subset( Afu, state=="Well" ), dur=NA ) )
> str( Dfu )
'data.frame':        542297 obs. of  11 variables:
 $ A     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ P     : num  1995 1995 1995 1995 1996 ...
 $ U     : num  0 0 0 0 0 0 0 0 1 1 ...
 $ dur   : num  0 0.2 0.4 0.6 0 0.2 0.4 0.6 0 0.2 ...
 $ sex   : Factor w/ 2 levels "M","F": 1 1 1 1 2 2 2 2 2 2 ...
 $ region: Factor w/ 8 levels "","Africa","America",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ state : Factor w/ 6 levels "Well","Dead",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ Y     : num  0.2 0.2 0.2 0.2 0.4 0.4 0.4 0.2 0.6 0.4 ...
 $ D.tb  : num  0 0 0 0 0 0 0 0 0 0 ...
 $ D.dm  : num  0 0 0 0 0 0 0 0 0 0 ...
 $ D.dd  : num  0 0 0 0 0 0 0 0 0 0 ...
```

The data frame `Dfu` now contains the correct number of person-years and transitions to TB and DM (but not the correct number of deaths from “Well”:

```
> round( addmargins( xtabs( Y ~ region + state, data=Dfu ) )/1000, 1 )
           state
region         Well     Dead       DM       TB   TB(DM)   DM(TB)      Sum
            73870.1      0.0   1600.5      0.0      0.0      0.0  75470.6
  Africa      306.6      0.0      7.1      0.0      0.0      0.0    313.7
  America     230.6      0.0      1.6      0.0      0.0      0.0    232.3
  Asia       1115.9      0.0     31.9      0.0      0.0      0.0   1147.8
  East_Euro   703.8      0.0     16.0      0.0      0.0      0.0    719.8
  Europe     1539.2      0.0     22.5      0.0      0.0      0.0   1561.7
  Oceania      31.2      0.0      0.1      0.0      0.0      0.0     31.3
  Other       137.7      0.0      3.5      0.0      0.0      0.0    141.3
  Sum       77935.2      0.0   1683.3      0.0      0.0      0.0  79618.4
```

```
> ftable( addmargins( xtabs( cbind(D.tb,D.dm,D.dd) ~ region + state,
+                            data=Dfu ),
+                     margin = 1:2 ),
+         row.vars=c(3,1) )
               state   Well   Dead     DM     TB TB(DM) DM(TB)    Sum
     region
D.tb                   3401      0    148      0      0      0   3549
     Africa            1242      0     17      0      0      0   1259
     America             24      0      0      0      0      0     24
     Asia               782      0     31      0      0      0    813
     East_Euro          161      0     10      0      0      0    171
     Europe             373      0     11      0      0      0    384
     Oceania              1      0      0      0      0      0      1
     Other              158      0      6      0      0      0    164
     Sum               6142      0    223      0      0      0   6365
D.dm                 295957      0      0      0      0      0 295957
     Africa            1250      0      0      0      0      0   1250
     America            361      0      0      0      0      0    361
     Asia                 5523      0      0      0      0      0   5523
     East_Euro            2823      0      0      0      0      0   2823
     Europe               4520      0      0      0      0      0   4520
     Oceania                36      0      0      0      0      0     36
     Other                 704      0      0      0      0      0    704
     Sum                311174      0      0      0      0      0 311174
D.dd                       234      0  75303      0      0      0  75537
     Africa                410      0     66      0      0      0    476
     America               274      0     23      0      0      0    297
     Asia                 1437      0    351      0      0      0   1788
     East_Euro            1589      0    401      0      0      0   1990
     Europe               2890      0    370      0      0      0   3260
     Oceania                34      0      3      0      0      0     37
     Other                 309      0     59      0      0      0    368
     Sum                  7177      0  76576      0      0      0  83753

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

2.3.3 Corrected boxes

We can now make boxes with the corrected no of person-years in the “Well” state, by getting the relevant data and doctoring the transition matrix appropriately (note that we put the number of transitions “Well” to “Dead” to 0 because we do not know this (yet!) for the entire population:

```
> load( file="../data/Afu.Rda" )
> addmargins( xtabs( D.tb ~ region + state, data=Dfu ) )

           state
region      Well Dead   DM   TB TB(DM) DM(TB)  Sum
            3401    0  148    0      0      0 3549
  Africa    1242    0   17    0      0      0 1259
  America     24    0    0    0      0      0   24
  Asia       782    0   31    0      0      0  813
  East_Euro  161    0   10    0      0      0  171
  Europe     373    0   11    0      0      0  384
  Oceania      1    0    0    0      0      0    1
  Other      158    0    6    0      0      0  164
  Sum       6142    0  223    0      0      0 6365

> formatC( at <- xtabs( cbind( D.tb, D.dm, D.dd, Y ) ~ state, data=Afu ),
+          format="f", digits=0, big.mark=",", preserve.width=NULL )

state    D.tb       D.dm       D.dd       Y
  Well        6,142    311,174      7,177 77,935,171
  Dead            0          0          0          0
  DM            326          0    136,088  2,629,386
  TB              0        228        258     46,319
  TB(DM)          0          0        126      1,638
  DM(TB)          0          0         44        944

> load( file="../data/xLx.Rda" )
> xLx <- subset( xLx, age-lex.dur <= 100 )
> ( tt <- tmat(        xLx            , Y=TRUE ) )

            Well   Dead      DM       TB   TB(DM)   DM(TB)
  Well   6529308   7185  311221  6142.00       NA       NA
  Dead        NA     NA      NA       NA       NA       NA
  DM          NA 136397 2629969       NA  326.000       NA
  TB          NA    258      NA 46372.17       NA 228.0000
  TB(DM)      NA    127      NA       NA 1639.064       NA
  DM(TB)      NA     44      NA       NA       NA 944.4052

> ( ti <- tmat( subset(xLx,region!=""), Y=TRUE ) )
```

```
          Well Dead       DM       TB   TB(DM)   DM(TB)
  Well   4065069 6951  15217.0  2741.00       NA       NA
  Dead        NA   NA       NA       NA       NA       NA
  DM          NA 1677 100507.4       NA 102.0000       NA
  TB          NA  122       NA 20249.66       NA 109.0000
  TB(DM)      NA   23       NA       NA 555.0198       NA
  DM(TB)      NA    2       NA       NA       NA 482.0424
```

```
> tt["Well","Well"] <- at["Well","Y"]
> ( td <- abs( tt - ti ) )
```

```
           Well   Dead      DM       TB   TB(DM)   DM(TB)
  Well   73870102    234  296004  3401.00       NA       NA
  Dead         NA     NA      NA       NA       NA       NA
  DM           NA 134720 2529461       NA  224.000       NA
  TB           NA    136      NA 26122.51       NA 119.0000
  TB(DM)       NA    104      NA       NA 1084.044       NA
  DM(TB)       NA     42      NA       NA       NA 462.3628
```

```
> td["Well","Dead"] <- 0
> tt["Well","Dead"] <- 0
```

With all PY transitions in place, we can show the final version of the transitions between states:

```
> aclr <- rep("black",9)
> aclr[5] <- "red"
> aclr[3] <- "forestgreen"
> tmpl <- function(){
+ par( mfrow=c(2,1) )
+ boxes.Lexis( td,
+              boxpos=list( x=c(10,90,10,50,50,90),
+                           y=c(65,35,35,90,10,65) ),
+              hmult=1.5, col.arr=aclr,
+              show=TRUE, scale.Y=1000, digits.R=2 )
+ text( 3, 95, "Danish born", adj=c(0,1), font=2, cex=1.5 )
+ boxes.Lexis( ti,
+              boxpos=list( x=c(10,90,10,50,50,90),
+                           y=c(65,35,35,90,10,65) ),
+              hmult=1.5, col.arr=aclr,
+              show=TRUE, scale.Y=1000 )
+ text( 3, 95, "Foreign born", adj=c(0,1), font=2, cex=1.5 )
+ }
> tmpl()
```

```
> pdf( "../graph/Fig1.pdf", height=14, width=10 )
> tmpl()
> dev.off()
```

```
null device
          1
```

```
> postscript( "../graph/Fig1.eps", height=14, width=10 )
> tmpl()
> dev.off()
```

```
null device
          1
```

```
> win.metafile( "../graph/Fig1.emf", height=14, width=10 )
> tmpl()
> dev.off()
```

```
null device
          1
```

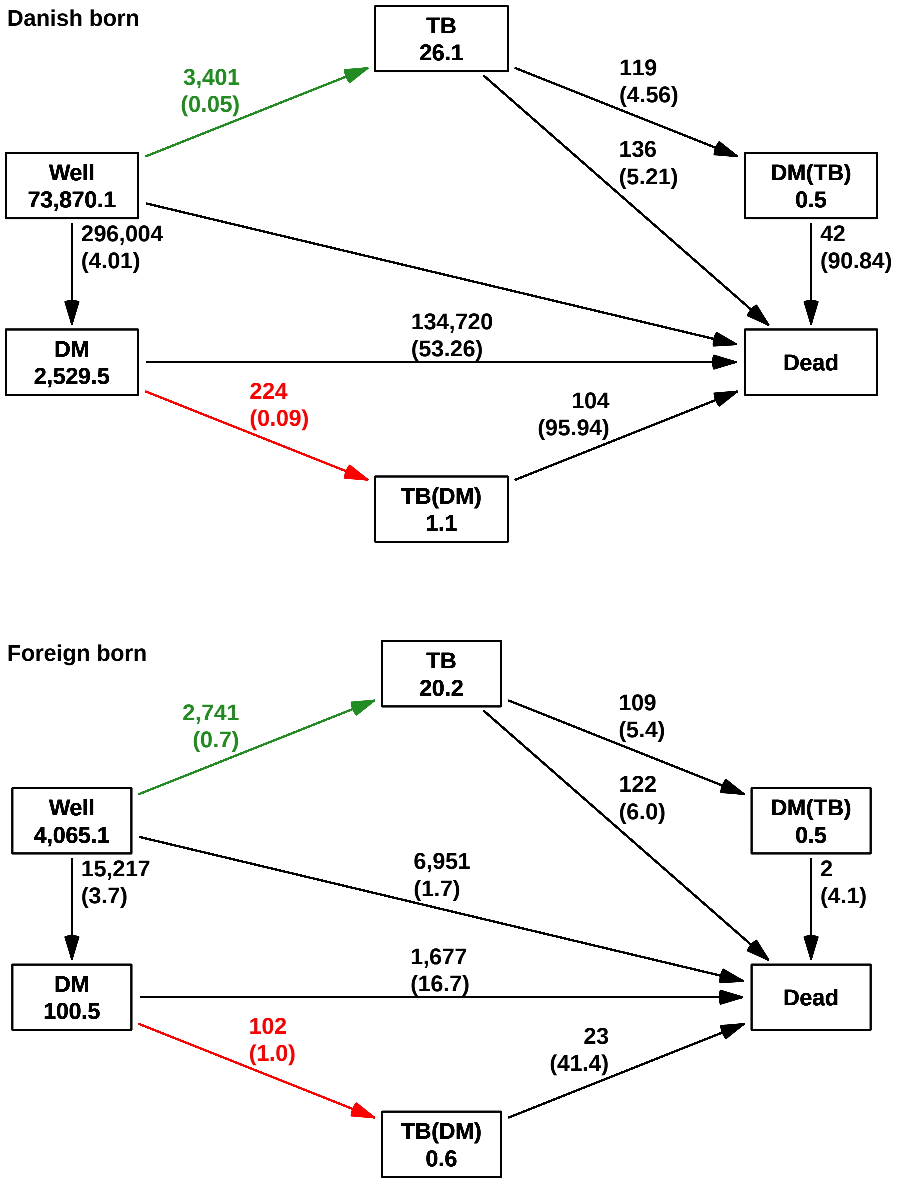

Figure 2.1: *Person-years (in 1000s) and number of transitions and -rates per 1000 PY for the DK population (top) and for the immigrants alone (bottom). Corresponds to figure 1 in the paper.*

Chapter 3

Analysis of incidence of TB

3.1 Overall TB incidence

The analysis using all diabetes patients regardless of date of diagnosis cannot include duration of diabetes, because date of diagnosis is only reliable for dates of diagnosis after 1995. Persons diagnosed before this do not have a reliable date of diagnosis recorded, and so diabetes duration is not reliably defined during their follow-up.

So first we reload the follow-up data:

```
> load( file="../data/Afu.Rda" )
> str( Afu )
'data.frame':        91922 obs. of  10 variables:
 $ sex   : Factor w/ 2 levels "M","F": 1 1 1 1 1 1 1 1 1 1 ...
 $ A     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ P     : num  1995 1995 1995 1995 1995 ...
 $ U     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ state : Factor w/ 6 levels "Well","Dead",..: 1 3 4 1 1 1 1 1 1 1 ...
 $ region: Factor w/ 8 levels "","Africa","America",..: 1 1 1 2 3 4 5 6 7 8 ...
 $ Y     : num  1.79e+04 8.06e-01 1.44e-01 4.53 2.05e+01 ...
 $ D.tb  : num  1 0 0 0 0 0 0 0 0 0 ...
 $ D.dm  : num  1 0 0 0 0 0 0 0 0 0 ...
 $ D.dd  : num  0 0 0 0 0 0 0 0 0 0 ...
```

We also need the splines package to model the effect of age properly, plus two little utilities to make life easier

```
> library( Epi )
> library( splines )
> source( "cnr.R" )
> cnr
function (xf, yf)
{
    cn <- par()$usr
    xf <- ifelse(xf > 1, xf/100, xf)
    yf <- ifelse(yf > 1, yf/100, yf)
    xx <- (1 - xf) * cn[1] + xf * cn[2]
    yy <- (1 - yf) * cn[3] + yf * cn[4]
    if (par()$xlog)
        xx <- 10^xx
    if (par()$ylog)
        yy <- 10^yy
    list(x = xx, y = yy)
}
> source( "rect.R" )
> rect
```

```
function (x1, y1, x2, y2, ...)
{
    if (is.list(x1)) {
        y1 <- x1$y
        x1 <- x1$x
    }
    if (length(x1) > 1 & length(y1) > 1)
        graphics::rect(x1[1], y1[1], x1[2], y1[2], ...)
    else graphics::rect(x1, y1, x2, y2, ...)
}
```

We want to have an overall picture of how TB incidence varies with ethnicity and age and how diabetes diagnosis influences this.

We will fit a Poisson model with terms in age and calendar time and categories of ethnicity (`Region`) and diabetes status (`state`). But since follow-up by age and period is classified in Lexis triangles we must recode age and period properly. The variable `U` is the indicator of the upper Lexis triangles, that is the part of an age×period group with the earliest data of birth, and hence the older age (A+2/3) and earlier time of observation. This recoding is done on the fly in subsetting the analysis data frame to the two states of interest:

```
> Atb <- transform( subset( Afu, state %in% c("Well","DM") ),
+                   state = factor(state),
+                  Region = Relevel( region, list(Asia=c(4,7),Other=c(3,5,6,8)), first=FALSE ),
+                      ax = A+(1+U)/3,
+                      px = P+(2-U)/3 )
> str( Atb )
'data.frame':        71434 obs. of  13 variables:
 $ sex   : Factor w/ 2 levels "M","F": 1 1 1 1 1 1 1 1 1 2 ...
 $ A     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ P     : num  1995 1995 1995 1995 1995 ...
 $ U     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ state : Factor w/ 2 levels "Well","DM": 1 2 1 1 1 1 1 1 1 1 ...
 $ region: Factor w/ 8 levels "","Africa","America",..: 1 1 2 3 4 5 6 7 8 1 ...
 $ Y     : num  1.79e+04 8.06e-01 4.53 2.05e+01 1.67e+01 ...
 $ D.tb  : num  1 0 0 0 0 0 0 0 0 0 ...
 $ D.dm  : num  1 0 0 0 0 0 0 0 0 0 ...
 $ D.dd  : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Region: Factor w/ 4 levels "","Africa","Asia",..: 1 1 2 4 3 4 4 3 4 1 ...
 $ ax    : num  0.333 0.333 0.333 0.333 0.333 ...
 $ px    : num  1996 1996 1996 1996 1996 ...
```

```
> levels( Atb$region )[1] <-
+ levels( Atb$Region )[1] <- "DK"
> ( atab <- addmargins( xtabs( D.tb ~ region + state, data=Atb ) ) )
           state
region      Well   DM  Sum
  DK        3401  224 3625
  Africa    1242   18 1260
  America     24    0   24
  Asia       782   51  833
  East_Euro  161   10  171
  Europe     373   16  389
  Oceania      1    0    1
  Other      158    7  165
  Sum       6142  326 6468
```

```
> ( aTab <- addmargins( xtabs( D.tb ~ Region + state, data=Atb ) ) )
        state
Region   Well   DM  Sum
  DK     3401  224 3625
  Africa 1242   18 1260
  Asia    783   51  834
  Other   716   33  749
  Sum    6142  326 6468
```

```
> atab <- rbind(atab,aTab[4,])[c(1,2,4,10,6,5,3,8,9),1:2]
> rownames( atab )[4] <- "Remain"
> atab
```

```
          Well  DM
DK        3401 224
Africa    1242  18
Asia       782  51
Remain     716  33
Europe     373  16
East_Euro  161  10
America     24   0
Other      158   7
Sum       6142 326
```

```
> save( atab, file="atab.Rda" )
```

We note that the number of TB cases is not overwhelming in the TB state:

```
> addmargins( xtabs( D.tb ~ region + state, data=Atb ) )
```

```
           state
region      Well   DM  Sum
  DK        3401  224 3625
  Africa    1242   18 1260
  America     24    0   24
  Asia       782   51  833
  East_Euro  161   10  171
  Europe     373   16  389
  Oceania      1    0    1
  Other      158    7  165
  Sum       6142  326 6468
```

so we would possibly be better off by a grouping of the region:

```
> Atb$Region <- Relevel( Atb$region, list(Asia=c(4,7),Other=c(3,5,6,8)), first=FALSE )
> formatC( ftable( addmargins( xtabs( cbind( D.tb, Y=Y/1000 ) ~ sex + Region + state,
+                                      data=Atb ),
+                              margin = 1:2 ),
+                  row.vars=2:1,
+                  col.vars=4:3 ),
+          format="f", digits=1, big.mark=",", pr="c" )
```

```
      [,1]       [,2]       [,3]       [,4]
 [1,] " 2,196.0" "   142.0" "36,457.2" " 1,289.5"
 [2,] " 1,205.0" "    82.0" "37,412.9" " 1,239.3"
 [3,] " 3,401.0" "   224.0" "73,870.1" " 2,528.9"
 [4,] "   668.0" "    10.0" "   163.4" "     4.6"
 [5,] "   574.0" "     8.0" "   143.2" "     3.7"
 [6,] " 1,242.0" "    18.0" "   306.6" "     8.3"
 [7,] "   345.0" "    30.0" "   549.1" "    19.3"
 [8,] "   438.0" "    21.0" "   598.0" "    20.9"
 [9,] "   783.0" "    51.0" " 1,147.1" "    40.2"
[10,] "   368.0" "    20.0" " 1,329.8" "    25.2"
[11,] "   348.0" "    13.0" " 1,281.6" "    26.8"
[12,] "   716.0" "    33.0" " 2,611.4" "    52.0"
[13,] " 3,577.0" "   202.0" "38,499.5" " 1,338.7"
[14,] " 2,565.0" "   124.0" "39,435.7" " 1,290.7"
[15,] " 6,142.0" "   326.0" "77,935.2" " 2,629.4"
attr(,"row.vars")
attr(,"row.vars")$Region
[1] "DK"     "Africa" "Asia"   "Other"  "Sum"

attr(,"row.vars")$sex
```

```
[1] "M"   "F"   "Sum"

attr(,"col.vars")
attr(,"col.vars")[[1]]
[1] "D.tb" "Y"

attr(,"col.vars")$state
[1] "Well" "DM"
```

. . . but even so, the number of TB cases among diabetes patients of African origin is below 20.

3.1.1 Analysis of DM effect on TB occurrence

First we set up the knots to use in the parametrization of the spline effects of age and period:

```
> nk <- 4
> ( a.kn <- with( Atb, quantile( rep(ax,D.tb), (1:nk-0.5)/nk ) ) )
   12.5%    37.5%    62.5%    87.5% 
19.33333 31.33333 43.66667 63.66667 
> nk <- 3
> ( p.kn <- with( Atb, quantile( rep(px,D.tb), (1:nk-0.5)/nk ) ) )

16.66667%       50% 83.33333% 
 1997.333  2001.667  2006.667 
```

We will also plot various curves etc. so we need a uniform color coding for the 4 groups:

```
> scol <- c("blue","red")
> names(scol) <- levels( Atb$sex )
> ecol <- c("black","orange","magenta","forestgreen")
> names(ecol) <- levels( Atb$Region )
> c( ecol, scol )
           DK        Africa          Asia         Other             M             F 
      "black"      "orange"     "magenta" "forestgreen"        "blue"         "red" 
> save( ecol, scol, file="../data/clrs.Rda" )
> par( mar=c(0,0,0,0) )
> plot(1:4,1:4,axes=F, xlab="", ylab="",xlim=c(0,5),ylim=c(0,5),type="n")
> text(rep(1,4),4:1, names(ecol), col=ecol, font=2, cex=2,adj=0 )
> text(rep(4,2),3:2, names(scol), col=scol, font=2, cex=2,adj=0 )
```

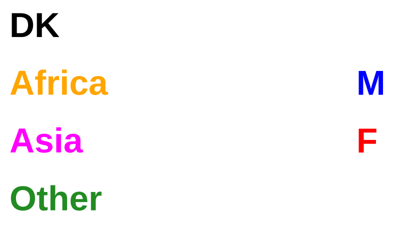

Figure 3.1: *Color coding used for the four geographic (ethnic) groups and for sex.*

Once all the paraphernalia has been set up, we fit three models for the TB-incidence.

3.1.1.1 Simple analysis

First we analyze the data without duration information, setting up a model with age, sex and presence of DM, and expand this by controlling for ethnicity (`region`) and by allowing an interaction with region.

This way we get two single estimates of RR_{DM}, one only controlled for age and sex, the other controlled also for ethnicity, and finally the interaction model provides estimates of RR_{DM} for each ethnicity. Note that we enter the persons-years in units of 10^5years, because we want to extract the estimated rates in those units too:

```
> m1 <- glm( D.tb ~ Ns( ax, kn=a.kn ) + Ns( px, kn=p.kn ) + sex + state,
+            offset = log(Y/10^5),
+            family = poisson,
+            data = Atb )
> m2 <- update( m1, . ~ . + Region )
> m3 <- update( m2, . ~ . - state + Region:state )
> anova( m3, m2, m1, test="Chisq" )
Analysis of Deviance Table

Model 1: D.tb ~ Ns(ax, kn = a.kn) + Ns(px, kn = p.kn) + sex + Region +
    state:Region
Model 2: D.tb ~ Ns(ax, kn = a.kn) + Ns(px, kn = p.kn) + sex + state +
    Region
Model 3: D.tb ~ Ns(ax, kn = a.kn) + Ns(px, kn = p.kn) + sex + state
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1     71420      17791
2     71423      17830 -3    -39.4 1.399e-08
3     71426      28982 -3 -11151.4 < 2.2e-16
```

```
> round( ci.exp( m1 ), 3 )
                   exp(Est.)   2.5%  97.5%
(Intercept)           10.875 10.315 11.466
Ns(ax, kn = a.kn)1     0.662  0.605  0.724
Ns(ax, kn = a.kn)2     1.558  1.439  1.688
Ns(ax, kn = a.kn)3     0.507  0.475  0.542
Ns(px, kn = p.kn)1     0.803  0.732  0.881
Ns(px, kn = p.kn)2     0.771  0.735  0.809
sexF                   0.706  0.672  0.742
stateDM                1.870  1.667  2.099
```

```
> round( ci.exp( m2 ), 3 )
                   exp(Est.)   2.5%  97.5%
(Intercept)            5.820  5.487  6.174
Ns(ax, kn = a.kn)1     1.070  0.977  1.172
Ns(ax, kn = a.kn)2     1.992  1.838  2.158
Ns(ax, kn = a.kn)3     1.052  0.979  1.130
Ns(px, kn = p.kn)1     0.581  0.529  0.638
Ns(px, kn = p.kn)2     0.646  0.616  0.678
sexF                   0.715  0.680  0.751
stateDM                1.598  1.425  1.793
RegionAfrica          87.720 81.946 93.902
RegionAsia            15.825 14.637 17.110
RegionOther            6.165  5.683  6.686
```

```
> round( ci.exp( m3 ), 3 )
                     exp(Est.)  2.5%  97.5%
(Intercept)              5.777 5.444  6.131
Ns(ax, kn = a.kn)1       1.082 0.988  1.186
Ns(ax, kn = a.kn)2       1.997 1.843  2.164
Ns(ax, kn = a.kn)3       1.053 0.980  1.131
Ns(px, kn = p.kn)1       0.581 0.529  0.638
Ns(px, kn = p.kn)2       0.647 0.616  0.679
sexF                     0.715 0.680  0.751
```

```
RegionAfrica            90.788 84.747 97.260
RegionAsia              15.818 14.596 17.143
RegionOther              6.115  5.627  6.645
RegionDK:stateDM         1.784  1.553  2.049
RegionAfrica:stateDM     0.521  0.327  0.831
RegionAsia:stateDM       1.762  1.325  2.344
RegionOther:stateDM      2.291  1.614  3.252
```

This provides pretty good evidence that not only does ethnicity influence the TB incidence, but the influence of DM on the TB incidence is different between ethnic groups.

It is of course also of interest *per se* to see how TB rates depend on age and on ethnicity, so from the last model (`m3`) we extract the age-specific incidence rates of TB among non-DM persons born in DK as well as the RR of TB in the non-DM population between each of the ethnic groups and the Danish born:

```
> n.pt <- 200
> a.pt <- seq(0,90,,n.pt)
> Ca <- Ns( a.pt, kn=a.kn )
> p.pt <- seq(1995,2010,,n.pt)
> Cp <- Ns( p.pt, kn=p.kn )
> p.ref <- 2005
> Cpr <- Ns( rep(p.ref,n.pt), kn=p.kn )
```

Once we have the contrast matrices (`Ca`, `Cp` and `Cpr`), we extract the age-specific

```
> m.eff <- ci.exp( m3, ctr.mat=cbind(1,Ca,Cpr), subset=c("Int","ax","px") )
> f.eff <- ci.exp( m3, ctr.mat=cbind(1,Ca,Cpr,1), subset=c("Int","ax","px","sex") )
> tmpl <- function(){
+ par( mar=c(3,3,1,1), mgp=c(3,1,0)/1.6 )
+ matplot( a.pt, m.eff, type="n", log="y",las=1, ylim=c(0.7,7),
+          xlab="Age at follow-up",
+          ylab="TB incidence rate per 100,000 PY in 2005")
+ abline( v=seq(0,90,5), h=c(5:15/10,2:10), col=gray(0.8) )
+ matlines( a.pt, cbind( m.eff, f.eff), lty=1, lwd=c(3,1,1),
+                 col=rep(scol,each=3) )
+ rect( cnr(c(0,10),c(85,100)), col="white", border=gray(0.8) )
+ text( cnr(rep(5,2),c(95,90)), levels(Atb$sex), col=scol, cex=1.1 )
+ box() }
> tmpl()
```

```
> p.rr <- ci.exp( m3, ctr.mat=Cp-Cpr, subset="px" )
> tmpl <- function() {
+ par( mar=c(3,3,1,1), mgp=c(3,1,0)/1.6 )
+ matplot( p.pt, p.rr, type="n", log="y",las=1, ylim=c(0.5,2),
+          xlab="Date of follow-up", ylab="RR of TB")
+ abline( v=1995:2010, h=c(1:15/10,2:10), col=gray(0.8) )
+ abline( h=1 )
+ matlines( p.pt, p.rr, lty=1, lwd=c(3,1,1), col="black" )
+ points( c(p.ref,p.ref), c(1,1), cex=1.3, pch=c(16,1), lwd=3,
+         col=c("white","black") ) }
> tmpl()
```

From figure3.2 it is seen that incidence rates increase to about age 25, after which they are pretty stable. From the figure 3.3 we see that the incidence rates seem to rater stably decreasing, and there is no significant curvature in the decline:

```
> m4 <- update( m3, . ~ . - Ns(px,kn=p.kn) + px )
> anova( m3, m4, test="Chisq" )
```

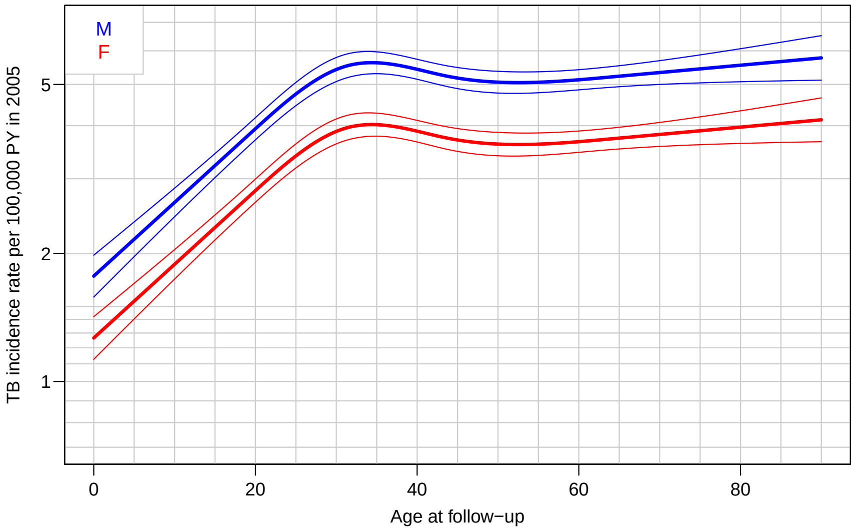

Figure 3.2: *Age-specific TB rates (2005) among non-diabetic men (blue) and women (red) born in Denmark. Also we see that under a proportionality assumption the RR comparing men to women is about 1.4.*

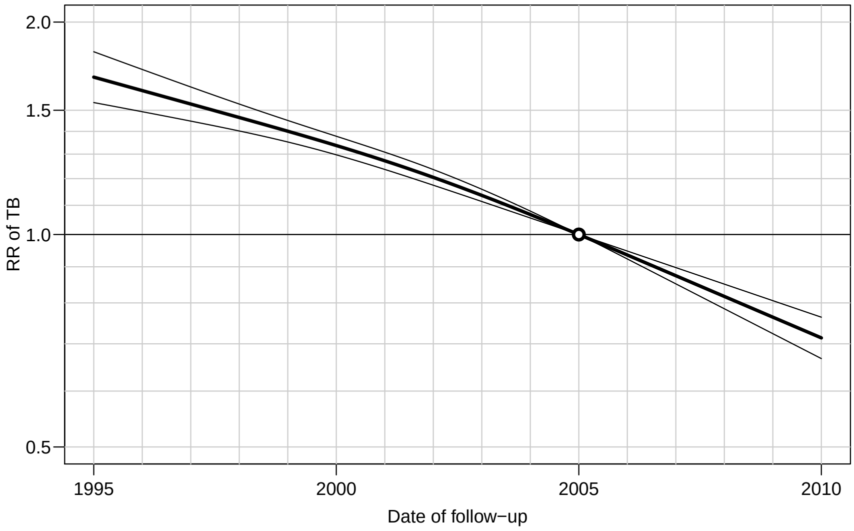

Figure 3.3: *TB-Rate-ratio by calendar time, relative to 2005.*

```
Analysis of Deviance Table

Model 1: D.tb ~ Ns(ax, kn = a.kn) + Ns(px, kn = p.kn) + sex + Region +
    state:Region
Model 2: D.tb ~ Ns(ax, kn = a.kn) + sex + Region + px + Region:state
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     71420      17791
2     71421      17794 -1  -3.0636  0.08006
```

```
> round((1-ci.exp( m4, subset="px" )[,c(1,3,2)])*100,2)
```

```
exp(Est.)      97.5%      2.5%
     5.50       4.96      6.05
```

On average the decrease in TB-rates is about 5.5% per year (95% c.i.: 4.9–6.0%) per year, corresponding to more than a halving over the 15-year period:

```
> round((ci.exp( m4, subset="px", ctr.mat=matrix(15,1,1) ))*100,2)

     exp(Est.)  2.5% 97.5%
[1,]     42.78 39.23 46.66
```

We also take a look at the RRs associated with the different groups:

```
> round( ci.exp( m3, subset="Region" ), 2 )

                      exp(Est.)  2.5% 97.5%
RegionAfrica              90.79 84.75 97.26
RegionAsia                15.82 14.60 17.14
RegionOther                6.11  5.63  6.64
RegionDK:stateDM           1.78  1.55  2.05
RegionAfrica:stateDM       0.52  0.33  0.83
RegionAsia:stateDM         1.76  1.33  2.34
RegionOther:stateDM        2.29  1.61  3.25
```

where we see a massive excess-incidence of TB among persons from Africa.

We then extract the estimates and plot the RRs of TB between persons with and without DM, both overall (i.e. only adjusted for age), adjusted for ethnicity, and with interaction with ethnicity:

```
> round( e1 <- ci.exp( m1, subset="state" ), 3 )

        exp(Est.)  2.5% 97.5%
stateDM      1.87 1.667 2.099

> round( e2 <- ci.exp( m2, subset="state" ), 3 )

        exp(Est.)  2.5% 97.5%
stateDM     1.598 1.425 1.793

> round( e3 <- ci.exp( m3, subset="state" ), 3 )

                      exp(Est.)  2.5% 97.5%
RegionDK:stateDM          1.784 1.553 2.049
RegionAfrica:stateDM      0.521 0.327 0.831
RegionAsia:stateDM        1.762 1.325 2.344
RegionOther:stateDM       2.291 1.614 3.252

> rownames( e3 ) <-
+ gsub( "Region","", gsub( ":stateDM", "", rownames( e3 ) ) )
> ee <- rbind( e1, e2, e3 )
> rownames( ee )[1:2] <- c("Raw","Region-adj")
> round( ee, 2 )

           exp(Est.) 2.5% 97.5%
Raw             1.87 1.67  2.10
Region-adj      1.60 1.42  1.79
DK              1.78 1.55  2.05
Africa          0.52 0.33  0.83
Asia            1.76 1.33  2.34
Other           2.29 1.61  3.25
```

We can plot these in a forest plot for comparison:

```
> par( mar=c(3,3,1,1), mgp=c(3,1,0)/1.6 )
> irr <- function(){
+ plotEst( ee[-(1:2),],
+          lwd=2, vref=1, cex=1.1, grid=c(3:15/10,2,2.5,3,4),
+          xtic=c(0.3,0.5,1,2,3,4),
+          xlab="TB RR: DM vs. non-DM", xlog=TRUE,
+          col=ecol, y=4:1 )}
> irr()
```

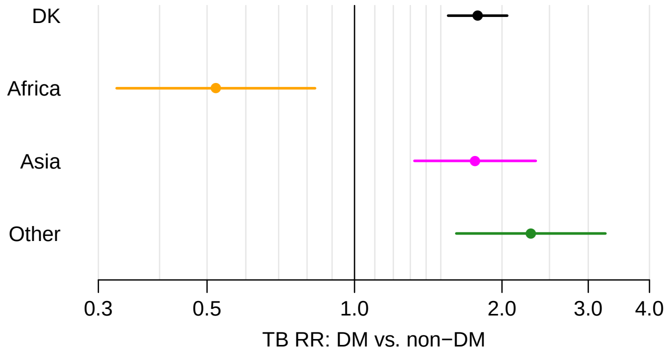

Figure 3.4: *Estimates of RR of TB associated with presence of DM. Separate estimates for each of the 4 ethnic subgroups (or rather groupings of country of birth)*

From figure 3.4 it is clear that the DM effect is much smaller among persons of African origin, but that is partly because the African born have a much higher overall incidence. So if we show the RRs relative to the DK-born non-DM persons we get the picture in figure 3.5:

```
> # Set up the relevant contrast matrix
> nr <- nlevels( Atb$Region )
> CRR <- diag( 2*nr )
> CRR[nr+1:nr,1:nr] <- diag(nr)
> CRR[1,] <- 0
> CRR <- CRR[,-1]
> rownames(CRR) <- t( outer( c("","DM "), levels(Atb$Region), paste, sep="" ) )
> CRR
          [,1] [,2] [,3] [,4] [,5] [,6] [,7]
DK           0    0    0    0    0    0    0
Africa       1    0    0    0    0    0    0
Asia         0    1    0    0    0    0    0
Other        0    0    1    0    0    0    0
DM DK        0    0    0    1    0    0    0
DM Africa    1    0    0    0    1    0    0
DM Asia      0    1    0    0    0    1    0
DM Other     0    0    1    0    0    0    1
```

```
> ci.exp( m3, subset="Region" )
```

```
                      exp(Est.)       2.5%      97.5%
RegionAfrica         90.7880591 84.7471186 97.2596097
RegionAsia           15.8183240 14.5960953 17.1428981
RegionOther           6.1147951  5.6271827  6.6446607
RegionDK:stateDM      1.7837503  1.5525751  2.0493469
RegionAfrica:stateDM  0.5210939  0.3269208  0.8305953
RegionAsia:stateDM    1.7624933  1.3254016  2.3437292
RegionOther:stateDM   2.2905812  1.6136029  3.2515822
```

```
> round( e3 <- ci.exp( m3, subset=c("Region"), ctr.mat=CRR ), 2 )
```

```
          exp(Est.)  2.5% 97.5%
DK             1.00  1.00  1.00
Africa        90.79 84.75 97.26
Asia          15.82 14.60 17.14
Other          6.11  5.63  6.64
DM DK          1.78  1.55  2.05
DM Africa     47.31 29.75 75.22
DM Asia       27.88 21.13 36.79
DM Other      14.01  9.94 19.74
```

```
> rownames( e3 )[nr+1:nr] <- NA
> par( mar=c(3,3,1,1), mgp=c(3,1,0)/1.6 )
> arr <- function(){
+ plotEst( e3, y=c(nr:1+0.1,nr:1-0.1), txtpos=rep(4:1,2),
+              lwd=2, vref=1, cex=1.1, grid=outer(1:9,10^(0:2)),
+              xtic=c(1,2,5,10,20,50,80,100),xlim=c(0.9,100),
+              xlab="TB RR vs. non-DM, DK born", xlog=TRUE,
+              col=c(rep(gray(0.5),4),ecol) )}
> arr()
```

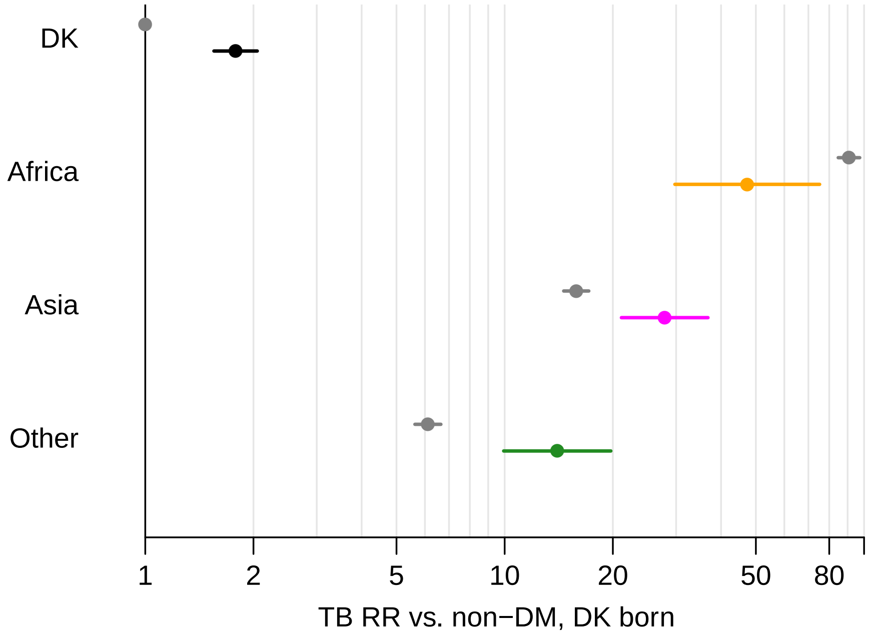

Figure 3.5: *Estimates of RR of TB relative to non-diabetic persons born in Denmark. The gray points represents estimates of RRs among persons without diabetes.*

```
> # PLot the two sets of RRs next to each other
> # First x-axis is from 0.3->4, the second from 1->100
> # The second is actually plotted from 10->1000
> rr2 <- function(){
+ par( mar=c(3,1,2,1), mgp=c(3,1,0)/1.6 )
+ plotEst( ee[-(1:2),], ylim=c(0,4.1),
+              xlim=c(0.3,1000), xtic=c(0.3,0.5,1,2,3,4),
+              lwd=2, vref=1, cex=1.1, #grid=outer(1:9,10^(0:2)),
+              xlog=TRUE, xlab="", grid=c(3:9/10,1.5,2:4),
+              col=ecol )
+ abline( v=c(1:9*10,1:9*100,1000,15,150), col=gray(0.9) )
+ abline( v=10 )
+ axis( side=1, at=c(1,2,5,c(1,2,5)*10,80,100)*10,
+           labels=c(1,2,5,c(1,2,5)*10,80,100) )
+ linesEst( e3*10, y=c(nr:1+0.1,nr:1-0.1), txtpos=rep(4:1,2),
+              lwd=2, col=c(rep(gray(0.5),4),ecol) )
+ mtext( "TB RR, DM vs. non-DM", side=1, at=1.0, line=2 )
+ mtext( "TB RR vs. DK born non-DM", side=1, at=100, line=2 )
+ mtext( c("a","b"), at=c(0.3,10)*1.1, side=3, line=1, font=2 )
+ }
> rr2()
```

```
> pdf("../graph/Fig2.pdf",height=2.2,width=7.5)
> rr2()
> dev.off()
> postscript("../graph/Fig2.eps",height=2.2,width=7.5)
> rr2()
> dev.off()
> # win.metafile("../graph/Fig2.emf",height=2.2,width=7.5)
> # rr2()
> # dev.off()
```

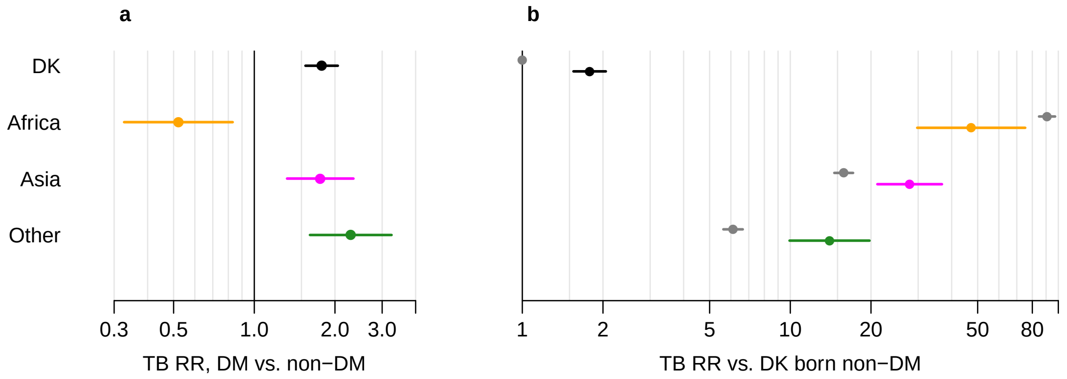

Figure 3.6: *TB RR* within *each region (place of birth) (left panel), and* across *regions using the Danish born without DM as reference (right panel). Corresponds to figure 2 in the paper.*

Figure 3.6 shows that TB rates are higher in "Other", even higher in "Asia" and highest in "Africa", both for rates among persons with and without diabetes. The TB rates among persons from Africa with DM just happens to be smaller than that among persons without.

3.1.2 Interactions

The three fitted models all build on a proportional hazards assumption, that is an assumption that the age-specific TB-rates are proportional between sexes, ethnic groups, and most boldly assumed, between persons with and without DM.

Thus we will look for the interaction effects between:

- age and sex
- age and region
- age and state
- period and region
- period and state

We first make formal likelihood-ratio-tests of these hypotheses, based on expanding the model `m3` (with a linear time-trend in incidence) successively with these interactions:

```
> mi     <- update( m3     , . ~ . - Ns(px,knots=p.kn) + I(px-2005) )
> mi.s   <- update( mi     , . ~ . +    sex:Ns(ax,knots=a.kn) )
> mi.sr  <- update( mi.s   , . ~ . + Region:Ns(ax,knots=a.kn) )
> mi.srs <- update( mi.sr  , . ~ . +  state:Ns(ax,knots=a.kn) )
> mi.Srs <- update( mi.srs , . ~ . +    sex:I(px-2005) )
> mi.SRs <- update( mi.Srs , . ~ . + Region:I(px-2005) )
> mi.SRS <- update( mi.SRs , . ~ . +  state:I(px-2005) )
> it <- as.matrix( anova( mi, mi.s  , mi.sr , mi.srs,
+                             mi.Srs, mi.SRs, mi.SRS,
+                         test="Chisq" ) )[-1,3:5]
> rownames( it ) <- c( "sex:age",
+                     "Region:age",
+                         "DM:age",
+                        "sex:per",
+                     "Region:per",
+                         "DM:per" )
> round( it, 3 )
           Df Deviance Pr(>Chi)
sex:age     3   77.677    0.000
Region:age  9  353.089    0.000
DM:age      3    4.811    0.186
sex:per     1    3.385    0.066
Region:per  3  154.664    0.000
DM:per      1    7.257    0.007
```

Thus it appears that there is little evidence of interactiosn between state and age and between sex and time, but that the interactions betwwen age and sex and region and time and region state are there:

```
> round( ci.exp( mi.SRS ), 3 )
                                exp(Est.)   2.5%   97.5%
(Intercept)                         3.245  3.007   3.503
Ns(ax, kn = a.kn)1                  2.430  2.107   2.803
Ns(ax, kn = a.kn)2                  3.260  2.872   3.699
Ns(ax, kn = a.kn)3                  1.704  1.530   1.898
sexF                                0.902  0.826   0.985
RegionAfrica                      105.399 94.103 118.052
RegionAsia                         16.904 14.750  19.373
RegionOther                         5.223  4.464   6.110
I(px - 2005)                        0.979  0.970   0.988
RegionDK:stateDM                    1.346  0.616   2.943
RegionAfrica:stateDM                0.597  0.247   1.441
RegionAsia:stateDM                  1.317  0.578   3.002
RegionOther:stateDM                 1.650  0.709   3.839
Ns(ax, kn = a.kn)1:sexF             0.479  0.399   0.576
Ns(ax, kn = a.kn)2:sexF             0.533  0.453   0.626
Ns(ax, kn = a.kn)3:sexF             0.622  0.545   0.709
Ns(ax, kn = a.kn)1:RegionAfrica     0.138  0.101   0.188
Ns(ax, kn = a.kn)2:RegionAfrica     0.340  0.267   0.434
Ns(ax, kn = a.kn)3:RegionAfrica     0.451  0.345   0.590
Ns(ax, kn = a.kn)1:RegionAsia       0.221  0.162   0.302
Ns(ax, kn = a.kn)2:RegionAsia       0.993  0.745   1.323
Ns(ax, kn = a.kn)3:RegionAsia       0.686  0.541   0.870
Ns(ax, kn = a.kn)1:RegionOther      0.385  0.284   0.524
Ns(ax, kn = a.kn)2:RegionOther      1.198  0.875   1.638
Ns(ax, kn = a.kn)3:RegionOther      0.848  0.666   1.081
Ns(ax, kn = a.kn)1:stateDM          1.463  0.790   2.711
Ns(ax, kn = a.kn)2:stateDM          1.377  0.295   6.436
Ns(ax, kn = a.kn)3:stateDM          1.142  0.708   1.842
sexF:I(px - 2005)                   0.989  0.978   1.001
RegionAfrica:I(px - 2005)           0.913  0.898   0.927
RegionAsia:I(px - 2005)             0.984  0.967   1.002
RegionOther:I(px - 2005)            0.933  0.916   0.951
I(px - 2005):stateDM                0.964  0.939   0.990
```

Thus the relevant model for description of the TB rates is one with the 4 interactions:

```
> m.int <- update( mi.SRS , . ~ . - state:Ns(ax,knots=a.kn)
+                                 - sex:I(px-2005) )
> round( ci.exp(m.int), 3 )
```

```
                                exp(Est.)   2.5%   97.5%
(Intercept)                         3.189  2.960   3.437
Ns(ax, kn = a.kn)1                  2.500  2.173   2.878
Ns(ax, kn = a.kn)2                  3.288  2.898   3.730
Ns(ax, kn = a.kn)3                  1.714  1.540   1.908
sexF                                0.935  0.864   1.011
RegionAfrica                      105.532 94.234 118.186
RegionAsia                         16.879 14.731  19.341
RegionOther                         5.230  4.471   6.118
I(px - 2005)                        0.974  0.967   0.982
RegionDK:stateDM                    1.565  1.343   1.824
RegionAfrica:stateDM                0.716  0.443   1.158
RegionAsia:stateDM                  1.562  1.148   2.126
RegionOther:stateDM                 1.971  1.355   2.867
Ns(ax, kn = a.kn)1:sexF             0.474  0.394   0.569
Ns(ax, kn = a.kn)2:sexF             0.528  0.450   0.621
Ns(ax, kn = a.kn)3:sexF             0.615  0.539   0.701
Ns(ax, kn = a.kn)1:RegionAfrica     0.136  0.100   0.186
Ns(ax, kn = a.kn)2:RegionAfrica     0.338  0.266   0.431
Ns(ax, kn = a.kn)3:RegionAfrica     0.450  0.344   0.588
Ns(ax, kn = a.kn)1:RegionAsia       0.225  0.165   0.306
Ns(ax, kn = a.kn)2:RegionAsia       0.989  0.742   1.317
Ns(ax, kn = a.kn)3:RegionAsia       0.684  0.539   0.867
Ns(ax, kn = a.kn)1:RegionOther      0.386  0.284   0.524
Ns(ax, kn = a.kn)2:RegionOther      1.188  0.869   1.625
Ns(ax, kn = a.kn)3:RegionOther      0.842  0.660   1.072
RegionAfrica:I(px - 2005)           0.912  0.898   0.927
RegionAsia:I(px - 2005)             0.984  0.966   1.001
RegionOther:I(px - 2005)            0.933  0.916   0.951
I(px - 2005):stateDM                0.965  0.940   0.990
```

This interaction model is now reported in a graph and a table:

- the age-specific TB-rates for the 4 ethnic groups in 2005, for men, separately for DM and non-DM persons.

- the age-specific M/F rate-ratio

- the annual change in TB-incidence rates for combinations of DM-status and region.

3.1.2.1 Age-interactions

First we derive the predicted rates for men in 2005:

```
> pp <- Atb[1:length(a.pt),c("sex","ax","px","Region","state")]
> str(pp)
'data.frame':        200 obs. of  5 variables:
 $ sex   : Factor w/ 2 levels "M","F": 1 1 1 1 1 1 1 1 1 2 ...
 $ ax    : num  0.333 0.333 0.333 0.333 0.333 ...
 $ px    : num  1996 1996 1996 1996 1996 ...
 $ Region: Factor w/ 4 levels "DK","Africa",..: 1 1 2 4 3 4 4 3 4 1 ...
 $ state : Factor w/ 2 levels "Well","DM": 1 2 1 1 1 1 1 1 1 1 ...

> pp <- transform( pp, sex = "M",
+                       ax = a.pt,
+                       px = 2005,
+                   Region = "DK",
+                    state = "Well",
+                        Y = 10^5 )
> head(pp)
```

```
  sex        ax   px Region state     Y
1   M 0.0000000 2005     DK  Well 1e+05
2   M 0.4522613 2005     DK  Well 1e+05
4   M 0.9045226 2005     DK  Well 1e+05
5   M 1.3567839 2005     DK  Well 1e+05
6   M 1.8090452 2005     DK  Well 1e+05
7   M 2.2613065 2005     DK  Well 1e+05
```

```
> ptb <- function( pp ) {
+ exp( do.call( cbind, predict( m.int,
+                               newdata=pp,
+                               type="link",
+                               se.fit=TRUE )[1:2] ) %*% ci.mat() ) }
> r.dk <- ptb( transform(pp,state="Well",Region="DK"    ) )
> d.dk <- ptb( transform(pp,state="DM"  ,Region="DK"    ) )
> r.af <- ptb( transform(pp,state="Well",Region="Africa") )
> d.af <- ptb( transform(pp,state="DM"  ,Region="Africa") )
> r.as <- ptb( transform(pp,state="Well",Region="Asia"  ) )
> d.as <- ptb( transform(pp,state="DM"  ,Region="Asia"  ) )
> r.ot <- ptb( transform(pp,state="Well",Region="Other" ) )
> d.ot <- ptb( transform(pp,state="DM"  ,Region="Other" ) )
> tmpf <- function(){
+ matplot( a.pt, cbind(r.dk,r.af,r.as,r.ot),
+          log="y", xlab="Age (years)", ylab="",
+          ylim=c(0.5,1800), xlim=c(10,85), las=1, yaxt="n",
+          col="transparent" )
+ axis( side=2, at=outer(c(1,2,5),10^(-1:4)),
+       labels=formatC( outer(c(1,2,5),10^(-1:4)), format="f", digits=1, drop=T ),
+       las=1 )
+ mtext( "TB incidence per 100,000 PY in 2005", side=2, line=2.5 )
+ abline( v=1:9*10, h=outer(c(1.5,1:9),10^c(-1:4)), col=gray(0.9) )
+ matlines( a.pt, cbind(r.dk,r.af,r.as,r.ot),
+          type="l", lty=1, lwd=c(4,1,1), col=rep(ecol,each=3) )
+ matlines( a.pt, cbind(d.dk,d.af,d.as,d.ot),
+          type="l", lty=rep(c("22","99"),c(1,2)), lwd=c(4,1,1),
+          lend=2, col=rep(ecol,each=3) )
+ text( cnr(2,101-c(4,1:3)*3), levels(Atb$Region), col=ecol, font=2, adj=c(0,1) )
+ abline( h=1 )
+ axis( side=4, at=c(5,7,10,15)/10, las=1 )
+ matlines( a.pt, ci.exp( m.int, subset="sex", ctr.mat=cbind(1,Ca) ),
+           type="l", lty=1, lwd=c(4,1,1), col="red" )
+ mtext( "Female/Male RR", side=4, at=1, line=2.0, col="red" )
+ box()
+ }
> par( mar=c(3,4,1,4), mgp=c(3,1,0)/1.6 )
> tmpf()
```

```
> pdf( "../graph/Fig3.pdf", width=8, height=8 )
> par( mar=c(3,4,1,4), mgp=c(3,1,0)/1.6 )
> tmpf()
> dev.off()
> postscript( "../graph/Fig3.eps", width=8, height=8 )
> par( mar=c(3,4,1,4), mgp=c(3,1,0)/1.6 )
> tmpf()
> dev.off()
> # win.metafile( "../graph/Fig3.emf", width=8, height=8 )
> # par( mar=c(3,4,1,4), mgp=c(3,1,0)/1.6 )
> # tmpf()
> # dev.off()
```

From figure 3.7 we see that for the non-Danish populations there is local peak at around 50 years with decreasing TB incidence rates after that, whereas the immigrant populations have a local peak at 35–40 and an increasing incidence by age. Moreover we see that up till about age 30 there is not much difference in TBN rates betwewern men and women, but women over 40 have substantially lower TB rates than man, with an RR of about 0.5–0.7.

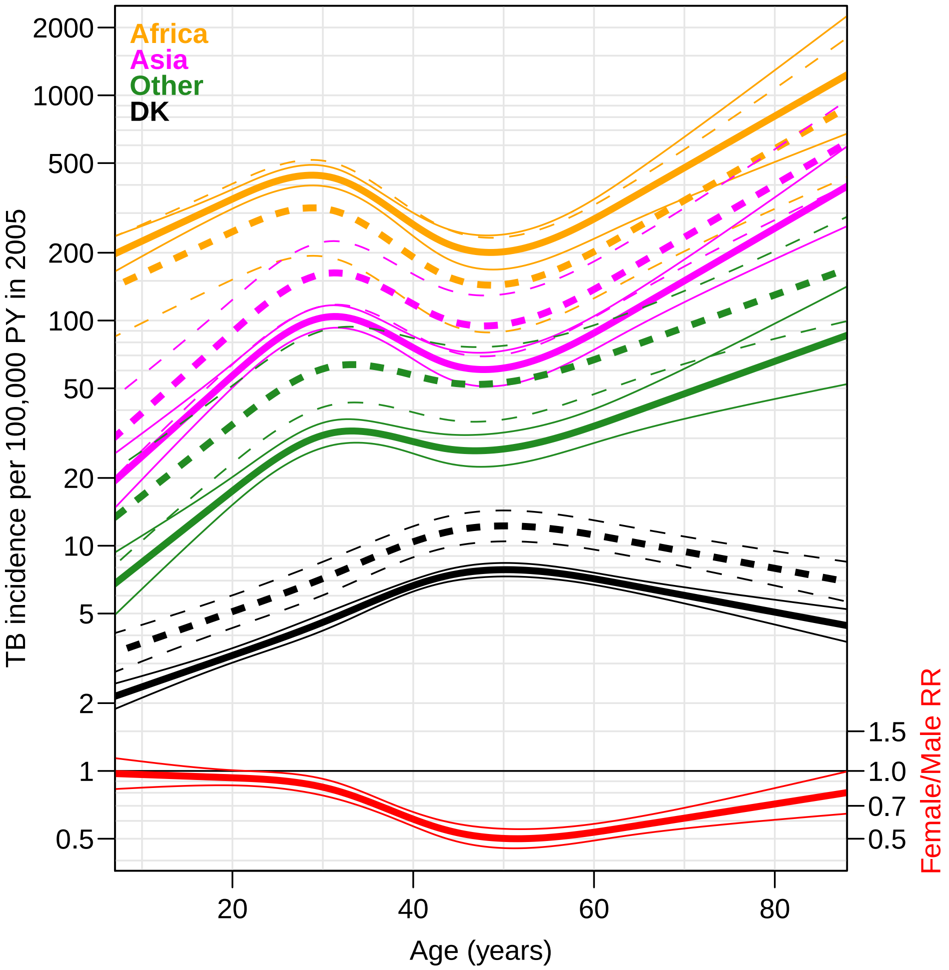

Figure 3.7: *Agw-specific rates of TB for men in 2005. Full lines are persons without DM, broken lines are for persons with diabetes. The full red line at the bottom is the female/male rate-ratio. All thin lines are 95% confidence intervals.*

3.1.2.2 Calendar time

In order to see how rates change by calendar time we estract the relevant parameters from the model:

```
> round( cfp <- ci.exp( m.int, subset="px" ), 3 )
                           exp(Est.)  2.5% 97.5%
I(px - 2005)                   0.974 0.967 0.982
RegionAfrica:I(px - 2005)      0.912 0.898 0.927
RegionAsia:I(px - 2005)        0.984 0.966 1.001
RegionOther:I(px - 2005)       0.933 0.916 0.951
I(px - 2005):stateDM           0.965 0.940 0.990
```

We want the annual change in TBN rates for any combination of region and DM status:

```
> rn <- outer( levels(Atb$state), levels(Atb$Region), paste )
> CM <- cbind( rep(1,8),
+              rep(c(0,1,0,0),each=2),
+              rep(c(0,0,1,0),each=2),
+              rep(c(0,0,0,1),each=2),
+              rep(0:1,4) )
```

```
> rownames( CM ) <- rn
> colnames( CM ) <- rownames( cfp )
> CM
            I(px - 2005) RegionAfrica:I(px - 2005) RegionAsia:I(px - 2005) RegionOther:I(px - 2005)
Well DK                1                         0                       0                        0
DM DK                  1                         0                       0                        0
Well Africa            1                         1                       0                        0
DM Africa              1                         1                       0                        0
Well Asia              1                         0                       1                        0
DM Asia                1                         0                       1                        0
Well Other             1                         0                       0                        1
DM Other               1                         0                       0                        1
            I(px - 2005):stateDM
Well DK                        0
DM DK                          1
Well Africa                    0
DM Africa                      1
Well Asia                      0
DM Asia                        1
Well Other                     0
DM Other                       1
> round( 100*(1-ci.exp(  m.int, subset="px", ctr.mat=CM ))[,c(1,3,2)], 1 )

            exp(Est.) 97.5% 2.5%
Well DK           2.6   1.8  3.3
DM DK             6.0   3.5  8.4
Well Africa      11.1   9.9 12.3
DM Africa        14.2  11.7 16.7
Well Asia         4.1   2.6  5.7
DM Asia           7.5   4.8 10.2
Well Other        9.1   7.5 10.6
DM Other         12.3   9.6 14.9
```

Chapter 4

Including duration of DM

4.1 TB incidence by DM duration

We now turn to the dataset which was split by diabetes duration, so first we reload the follow-up data:

```
> load( file="../data/Dfu.Rda" )
> str( Dfu )
'data.frame':        542297 obs. of  11 variables:
 $ A     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ P     : num  1995 1995 1995 1995 1996 ...
 $ U     : num  0 0 0 0 0 0 0 0 1 1 ...
 $ dur   : num  0 0.2 0.4 0.6 0 0.2 0.4 0.6 0 0.2 ...
 $ sex   : Factor w/ 2 levels "M","F": 1 1 1 1 2 2 2 2 2 2 ...
 $ region: Factor w/ 8 levels "","Africa","America",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ state : Factor w/ 6 levels "Well","Dead",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ Y     : num  0.2 0.2 0.2 0.2 0.4 0.4 0.4 0.2 0.6 0.4 ...
 $ D.tb  : num  0 0 0 0 0 0 0 0 0 0 ...
 $ D.dm  : num  0 0 0 0 0 0 0 0 0 0 ...
 $ D.dd  : num  0 0 0 0 0 0 0 0 0 0 ...
> with( Dfu, ftable( state, durOK=!is.na(dur), D.tb, row.vars=1:2 ) )
             D.tb      0      1      2      3      4      5      6      7
state  durOK
Well   FALSE       38843   3422    827    239     59     15      4      2
       TRUE            0      0      0      0      0      0      0      0
Dead   FALSE           0      0      0      0      0      0      0      0
       TRUE            0      0      0      0      0      0      0      0
DM     FALSE           0      0      0      0      0      0      0      0
       TRUE       498663    223      0      0      0      0      0      0
TB     FALSE           0      0      0      0      0      0      0      0
       TRUE            0      0      0      0      0      0      0      0
TB(DM) FALSE           0      0      0      0      0      0      0      0
       TRUE            0      0      0      0      0      0      0      0
DM(TB) FALSE           0      0      0      0      0      0      0      0
       TRUE            0      0      0      0      0      0      0      0
```

When we model the effect of duration of diabetes on TB incidence, we must provide a valid value for this variable for the non-diabetics; we will use 0. As for the entire dataset above, we also need to define the midpoints of the follow-up in the Lexis triangles:

```
> Dtb <- transform( subset( Dfu, state %in% c("Well","DM") ),
+                   state = factor(state),
+                  Region = Relevel( region, list(Asia=c(4,7),Other=c(3,5,6,8)), first=FALSE ),
+                      ax = A+(1+U)/3,
+                      px = P+(2-U)/3,
+                     dur = pmax( Dfu$dur, 0, na.rm=TRUE ) )
```

With this recoding we now can make the same tabulation as for the large dataset, and compare the two:

```
> ( dtab <- addmargins( xtabs( D.tb ~ region + state, data=Dtb ) ) )
           state
region      Well   DM  Sum
            3401  148 3549
  Africa    1242   17 1259
  America     24    0   24
  Asia       782   31  813
  East_Euro  161   10  171
  Europe     373   11  384
  Oceania      1    0    1
  Other      158    6  164
  Sum       6142  223 6365

> ( dTab <- addmargins( xtabs( D.tb ~ Region + state, data=Dtb ) ) )

        state
Region   Well   DM  Sum
         3401  148 3549
  Africa 1242   17 1259
  Asia    783   31  814
  Other   716   27  743
  Sum    6142  223 6365

> dtab <- rbind(dtab,dTab[4,])[c(1,2,4,10,6,5,3,8,9),1:2]
> rownames( dtab )[c(1,4)] <- c("DK","Remain")
> colnames( dtab )[2] <- "DM(dur)"
> load( file="atab.Rda" )
> cbind( atab, dtab )

          Well  DM Well DM(dur)
DK        3401 224 3401     148
Africa    1242  18 1242      17
Asia       782  51  782      31
Remain     716  33  716      27
Europe     373  16  373      11
East_Euro  161  10  161      10
America     24   0   24       0
Other      158   7  158       6
Sum       6142 326 6142     223

> ( tt <- cbind( atab, dtab )[,-3] )

          Well  DM DM(dur)
DK        3401 224     148
Africa    1242  18      17
Asia       782  51      31
Remain     716  33      27
Europe     373  16      11
East_Euro  161  10      10
America     24   0       0
Other      158   7       6
Sum       6142 326     223

> cbind( tt, round( sweep( tt, 2, tt[9,]/100, "/" ), 1 ) )[,c(1,4,2,5,3,6)]

          Well  Well  DM    DM DM(dur) DM(dur)
DK        3401  55.4 224  68.7     148    66.4
Africa    1242  20.2  18   5.5      17     7.6
Asia       782  12.7  51  15.6      31    13.9
Remain     716  11.7  33  10.1      27    12.1
Europe     373   6.1  16   4.9      11     4.9
East_Euro  161   2.6  10   3.1      10     4.5
America     24   0.4   0   0.0       0     0.0
Other      158   2.6   7   2.1       6     2.7
Sum       6142 100.0 326 100.0     223   100.0
```

As seen we have the same number of TB events in the “well” state for both types of analyses.

In the modeling we must make sure that persons with a value of 0 for duration and *not* in the “DM” state is modeled with 0 contribution from the duration term. This is achieved by letting the spline term in duration have its left boundary knot equal to 0:

```
> nk <- 4
> p.dur <- seq(0,16,,100)
> ( d.kn <- with( subset(Dtb,state=="DM"),
+                  c( 0,
+                     quantile( rep(dur,D.tb),
+                               1:nk/(nk+0.5) ) ) ) )
          22.22222% 44.44444% 66.66667% 88.88889%
0.0000000 0.6666667 2.2000000 3.8000000 8.2000000
> ( dmd <- with( subset(Dtb,state=="DM"), median( rep(dur,D.tb) ) ) )
[1] 2.6
```

So we see that half of the TB cases among diabetes patients occur before 2.6 years of diabetes duration.

```
> matplot( p.dur, Ns(p.dur,knots=d.kn), type="l", lwd=3, lty=1 )
> rug( d.kn, lwd=3 )
```

We also need the spline knots for age:

```
> nk <- 4
> ( a.kn <- with( Dtb, quantile( rep(ax,D.tb), (1:nk-0.5)/nk ) ) )
   12.5%    37.5%    62.5%    87.5%
18.66667 31.33333 43.33333 63.33333
> nk <- 3
> ( p.kn <- with( Dtb, quantile( rep(px,D.tb), (1:nk-0.5)/nk ) ) )
16.66667%       50% 83.33333%
 1997.333  2001.667  2006.667
```

In the light of the models fitted for the entire dataset, we must include main effects of age (a spline), diabetes presence, ethnicity (*i.e.* region of birth, `region`).

Strictly speaking we should also include interactions between age and ethnicity and sex, but we will omit this in the first instance, because the data base (in terms of TB cases) is somewhat smaller, and we must expand the model with the duration term.

So we fit the model where the only interaction we include is the state by ethnicity interaction. This means that when showing the duration effect, we will show it for a specific group (DK) and that the effect for the other groups is just offset a constant over the entire duration spectrum.

```
> md <- glm( D.tb ~ Ns( ax , kn=a.kn ) + sex + Region*state +
+                   Ns( dur, kn=d.kn ) + Region:I(px-2005),
+            offset = log(Y/10^5),
+            family = poisson,
+            data = Dtb )
> summary( md )
Call:
glm(formula = D.tb ~ Ns(ax, kn = a.kn) + sex + Region * state +
    Ns(dur, kn = d.kn) + Region:I(px - 2005), family = poisson,
    data = Dtb, offset = log(Y/10^5))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
```

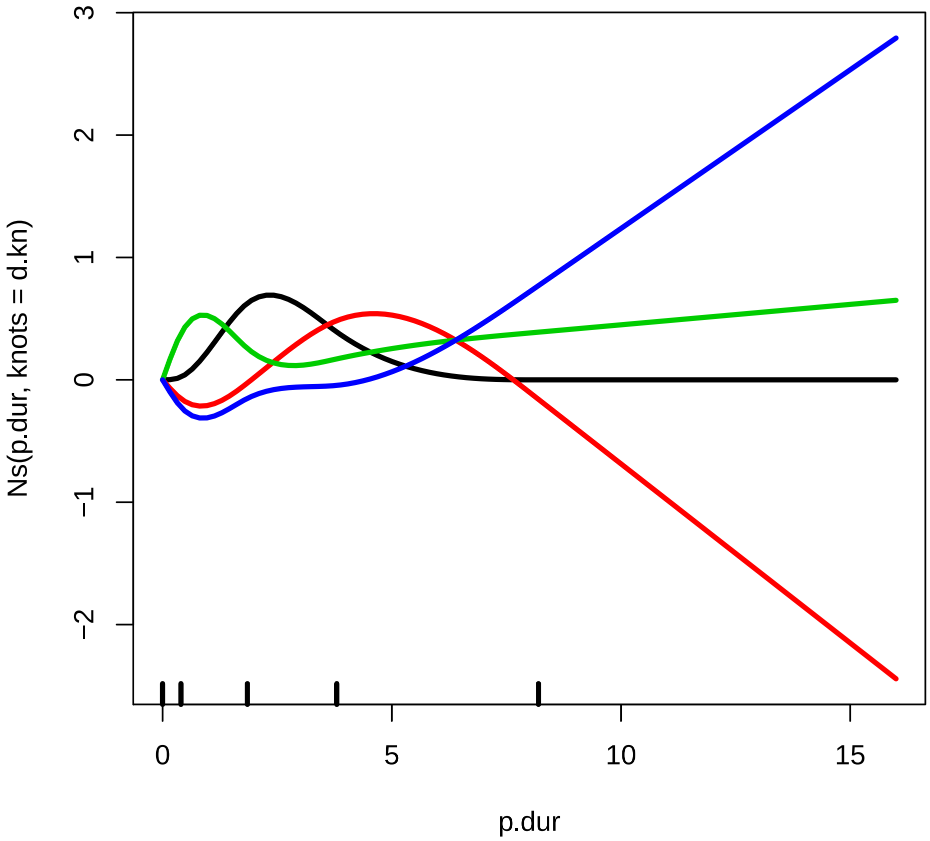

Figure 4.1: *Illustration of the spline basis used for duration; the point here is the fact that all components are 0 in 0, and thus the duration terms contributes nothing to the model for the rates among the non-DM persons.*

```
-2.0270  -0.0359  -0.0216  -0.0140   5.2717

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                1.390743   0.030642  45.386  < 2e-16
Ns(ax, kn = a.kn)1         0.109201   0.046954   2.326 0.020034
Ns(ax, kn = a.kn)2         0.756718   0.042596  17.765  < 2e-16
Ns(ax, kn = a.kn)3         0.081297   0.036689   2.216 0.026702
sexF                      -0.325926   0.025522 -12.770  < 2e-16
RegionAfrica               4.190066   0.045083  92.941  < 2e-16
RegionAsia                 2.697510   0.046343  58.208  < 2e-16
RegionOther                1.605198   0.052062  30.833  < 2e-16
stateDM                    1.313315   0.202245   6.494 8.38e-11
Ns(dur, kn = d.kn)1       -0.266074   0.278446  -0.956 0.339289
Ns(dur, kn = d.kn)2       -0.702355   0.244413  -2.874 0.004058
Ns(dur, kn = d.kn)3       -2.013582   0.547893  -3.675 0.000238
Ns(dur, kn = d.kn)4       -0.403936   0.181622  -2.224 0.026145
RegionAfrica:stateDM      -0.977598   0.259349  -3.769 0.000164
RegionAsia:stateDM        -0.307869   0.202701  -1.519 0.128803
RegionOther:stateDM        0.283047   0.214577   1.319 0.187139
Region:I(px - 2005)       -0.026973   0.003931  -6.862 6.79e-12
RegionAfrica:I(px - 2005) -0.129812   0.006981 -18.596  < 2e-16
```

```
RegionAsia:I(px - 2005)   -0.044031   0.008311  -5.298 1.17e-07
RegionOther:I(px - 2005)  -0.096100   0.008736 -11.000  < 2e-16

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 31251  on 542296  degrees of freedom
Residual deviance: 18543  on 542277  degrees of freedom
AIC: 29017

Number of Fisher Scoring iterations: 10
```

```
> round( ci.exp( md ), 3 )
                          exp(Est.)   2.5%  97.5%
(Intercept)                   4.018  3.784  4.267
Ns(ax, kn = a.kn)1            1.115  1.017  1.223
Ns(ax, kn = a.kn)2            2.131  1.961  2.317
Ns(ax, kn = a.kn)3            1.085  1.009  1.166
sexF                          0.722  0.687  0.759
RegionAfrica                 66.027 60.443 72.127
RegionAsia                   14.843 13.554 16.254
RegionOther                   4.979  4.496  5.514
stateDM                       3.718  2.502  5.527
Ns(dur, kn = d.kn)1           0.766  0.444  1.323
Ns(dur, kn = d.kn)2           0.495  0.307  0.800
Ns(dur, kn = d.kn)3           0.134  0.046  0.391
Ns(dur, kn = d.kn)4           0.668  0.468  0.953
RegionAfrica:stateDM          0.376  0.226  0.625
RegionAsia:stateDM            0.735  0.494  1.094
RegionOther:stateDM           1.327  0.872  2.021
Region:I(px - 2005)           0.973  0.966  0.981
RegionAfrica:I(px - 2005)     0.878  0.866  0.890
RegionAsia:I(px - 2005)       0.957  0.941  0.973
RegionOther:I(px - 2005)      0.908  0.893  0.924
```

Having fitted the model we can now extract the duration effects; we need the duration term plus the effect of DM:

```
> load( file="../data/clrs.Rda" )
> d.pt <- seq(0,15,,200)
> Cd <- Ns( d.pt, kn=d.kn )
> round( ci.exp( md, subset=c("dur","DM") ), 3 )
                     exp(Est.)  2.5% 97.5%
Ns(dur, kn = d.kn)1      0.766 0.444 1.323
Ns(dur, kn = d.kn)2      0.495 0.307 0.800
Ns(dur, kn = d.kn)3      0.134 0.046 0.391
Ns(dur, kn = d.kn)4      0.668 0.468 0.953
stateDM                  3.718 2.502 5.527
RegionAfrica:stateDM     0.376 0.226 0.625
RegionAsia:stateDM       0.735 0.494 1.094
RegionOther:stateDM      1.327 0.872 2.021
```

```
> d.eff <- ci.exp( md, subset=c("dur","DM"), ctr.mat=cbind(Cd,1,0,0,0) )
> d.afr <- ci.exp( md, subset=c("dur","DM"), ctr.mat=cbind(Cd,1,1,0,0) )
> d.asi <- ci.exp( md, subset=c("dur","DM"), ctr.mat=cbind(Cd,1,0,1,0) )
> d.oth <- ci.exp( md, subset=c("dur","DM"), ctr.mat=cbind(Cd,1,0,0,1) )
> tmpl <- function() {
+ par( mar=c(3,3,1,1), mgp=c(3,1,0)/1.6 )
+ matplot( d.pt, d.eff, type="n",
+          log="y",las=1, ylim=c(0.2,5),
+          xlab="DM duration (years)", ylab="RR of TB versus non-DM persons")
+ abline( v=seq(0,16,2), h=c(5:15/10,2:10), col=gray(0.8) )
+ abline( h= 1 )
+ matlines( d.pt, cbind(d.eff,d.afr,d.asi,d.oth),
+           type="l", lty=1, lwd=c(5,2,2),
+           col=rep(ecol,each=3) )
+ rect( cnr(c(80,100),c(80,100)), col="white", border=gray(0.8) )
+ cc <- cnr(82,97)
+ text( cc$x, cc$y*0.85^(0:3), names(ecol)[c(4,1,3,2)], col=ecol[c(4,1,3,2)], adj=0 )
+ box() }
> tmpl()
```

```
> win.metafile( "../graph/dur1.emf", width=8, height=6, pointsize=14)
> tmpl()
> dev.off()
```

```
null device
          1
```

It is clear from figure ?? that the duration effect is over-modeled with 5 knots. Moreover, it is difficult to believe that the RR is proportional between the 4 groups, but this is a bit more tricky to investigate precisely because of the rather limited number of tuberculosis cases among diabetes patients in these groups:

```
> xtabs( D.tb ~ Region + state, data=Dtb )
```

```
        state
Region   Well   DM
         3401  148
  Africa 1242   17
  Asia    783   31
  Other   716   27
```

First, we re-do the model with only 3 knots:

```
> nk <- 3
> d.kn <- with( subset(Dfu,state=="DM"),
+               c( 0,
+                  quantile( rep(dur,D.tb),
+                            1:nk/(nk+0.5) ) ) )
> md <- update( md, . ~ . )
> Cd <- Ns( d.pt, kn=d.kn )
> d.eff <- ci.exp( md, subset=c("dur","DM"), ctr.mat=cbind(Cd,1,0,0,0) )
> d.afr <- ci.exp( md, subset=c("dur","DM"), ctr.mat=cbind(Cd,1,1,0,0) )
> d.asi <- ci.exp( md, subset=c("dur","DM"), ctr.mat=cbind(Cd,1,0,1,0) )
> d.oth <- ci.exp( md, subset=c("dur","DM"), ctr.mat=cbind(Cd,1,0,0,1) )
> tmpl <- function(){
+ par( mar=c(3,3,1,1), mgp=c(3,1,0)/1.6 )
+ matplot( d.pt, d.eff, type="n",
+          log="y",las=1, ylim=c(0.2,5), xlim=c(0,13),
+          xlab="DM duration (years)", ylab="RR of TB versus non-DM persons")
+ abline( v=seq(0,16,1), h=c(5:10/10,1.5,2:10), col=gray(0.8) )
+ abline( h= 1 )
+ matlines( d.pt, cbind(d.eff,d.afr,d.asi,d.oth),
+           type="l", lty=1, lwd=c(5,2,2),
+           col=rep(ecol,each=3) )
+ # rect( cnr(c(86,100),c(80,100)), col="white", border=gray(0.8) )
+ cc <- cnr(3,3)
+ text( cc$x, cc$y*1.15^(0:3), names(ecol)[c(2,3,1,4)], col=ecol[c(2,3,1,4)], adj=c(0,0) )
+ box() }
> tmpl()
```

It is however of interest to see if there is any interaction between duration and ethnic group, but because of the limited data, we only include a linear-duration by group interaction, that is the only extra variation we introduce between the curves in figure ?? is an individual tilt to them, the basic shape is assumed to be the same:

```
> gc()
```

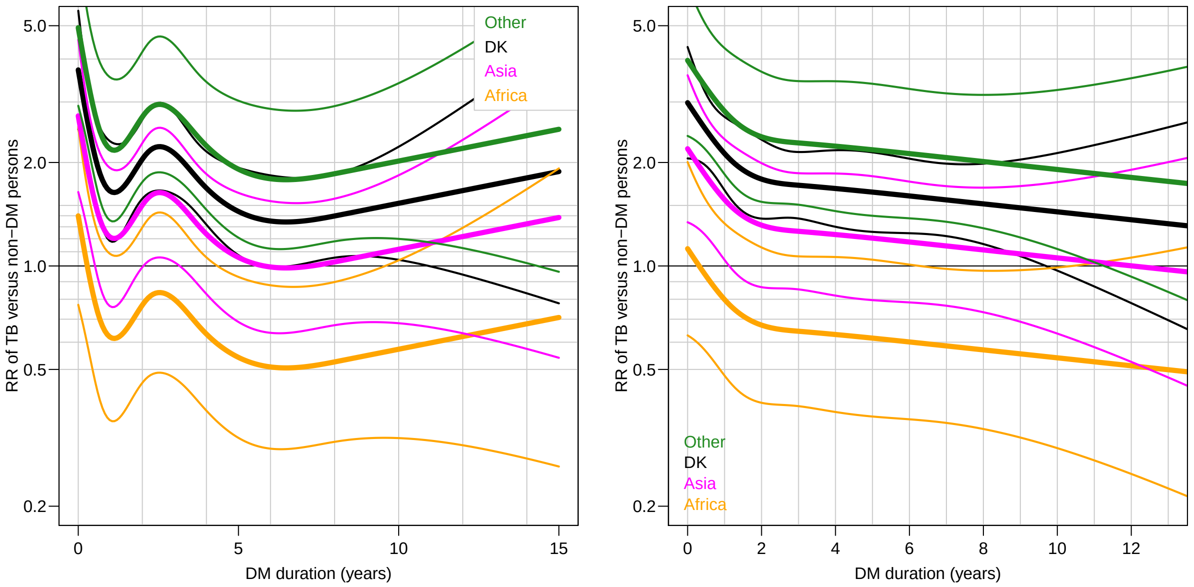

Figure 4.2: *Duration-specific RRs of TB occurrence for the 4 groups, under the assumption that all effects are proportional. 5 (left) and 4 (right) knots used for the duration effects.*

```
           used  (Mb) gc trigger  (Mb)  max used   (Mb)
Ncells   829835  22.2    1368491  36.6    928161   24.8
Vcells 43741009 333.8  124884103 952.8 155929817 1189.7
```

```
> mdi <- update( md, . ~ . + Region:dur )
> round( ci.exp( mdi ), 3 )
```

```
                          exp(Est.)   2.5%  97.5%
(Intercept)                   4.017  3.783  4.265
Ns(ax, kn = a.kn)1            1.115  1.017  1.223
Ns(ax, kn = a.kn)2            2.131  1.961  2.317
Ns(ax, kn = a.kn)3            1.085  1.010  1.166
sexF                          0.722  0.687  0.759
RegionAfrica                 66.060 60.470 72.166
RegionAsia                   14.845 13.555 16.257
RegionOther                   4.986  4.502  5.522
stateDM                       2.900  1.962  4.287
Ns(dur, kn = d.kn)1           0.594  0.325  1.086
Ns(dur, kn = d.kn)2           0.258  0.077  0.867
Ns(dur, kn = d.kn)3           0.512  0.208  1.258
RegionAfrica:stateDM          0.406  0.195  0.847
RegionAsia:stateDM            0.736  0.407  1.330
RegionOther:stateDM           1.549  0.853  2.816
Region:I(px - 2005)           0.973  0.966  0.981
RegionAfrica:I(px - 2005)     0.878  0.866  0.890
RegionAsia:I(px - 2005)       0.957  0.941  0.973
RegionOther:I(px - 2005)      0.909  0.893  0.924
Region:dur                    1.054  0.906  1.226
RegionAfrica:dur              1.028  0.822  1.285
RegionAsia:dur                1.053  0.877  1.264
RegionOther:dur               1.000  1.000  1.000
```

```
> 1-pchisq( md$deviance - mdi$deviance,
+           md$df.res   - mdi$df.res )
```

```
[1] 0.910785
```

So we see there is no interaction of any significance, which is also pretty clear from figure **??** when looking at the confidence intervals for the lines.

```
> d.eff <- ci.exp( mdi, subset=c("dur","DM"), ctr.mat=cbind(Cd,d.pt, 0  ,0,0,1,0,0,0) )
> d.afr <- ci.exp( mdi, subset=c("dur","DM"), ctr.mat=cbind(Cd,d.pt,d.pt,0,0,1,1,0,0) )
> d.asi <- ci.exp( mdi, subset=c("dur","DM"), ctr.mat=cbind(Cd,d.pt,0,d.pt,0,1,0,1,0) )
> d.oth <- ci.exp( mdi, subset=c("dur","DM"), ctr.mat=cbind(Cd,d.pt,0,0,d.pt,1,0,0,1) )
> par( mar=c(3,3,1,1), mgp=c(3,1,0)/1.6 )
> matplot( d.pt, d.eff, type="n",
+          log="y",las=1, ylim=c(0.2,5), xlim=c(0,13),
+          xlab="DM duration (years)", ylab="RR of TB versus non-DM persons")
> abline( v=seq(0,16,2), h=c(5:10/10,1.5,2:10), col=gray(0.8) )
> abline( h= 1 )
> matlines( d.pt, cbind(d.eff,d.afr,d.asi,d.oth),
+           type="l", lty=1, lwd=c(3,1,1),
+           col=rep(ecol,each=3) )
> box()
```

This means that a sensible summary of the DM-duration effect is as in model `md`, where there is a single function to describe the DM-duration effect, and a fixed (not varying by DM duration) effect to describe the effect of ethnicity.

4.1.1 Calendar time interaction

In the analysis of the dataset without duration information, we saw a substantial interaction with calendar time, that is we saw that the difference in TB incidence rates between DM and non-DM persons diminished. It would therefore be of interest to expand the model `md` with a calendar time term and an interaction between duration and calendar time. The model fitted were:

```
> md <- glm( D.tb ~ Ns( ax , kn=a.kn ) + sex + Region*state +
+                   Ns( dur, kn=d.kn ),
+            offset = log(Y/10^5),
+            family = poisson,
+            data = Dtb )
```

and we now expand it with a duration by calendar time interaction:

```
> md.dev <- md$deviance
> md.df  <- md$df.res
> rm( md )
> gc()
           used  (Mb) gc trigger   (Mb)  max used   (Mb)
Ncells   830146  22.2    1368491   36.6    928161   24.8
Vcells 46585357 355.5  134320633 1024.8 167879470 1280.9
```

```
> mdp <- glm( D.tb ~ Ns( ax , kn=a.kn ) + sex + Region*state +
+                                 state:Ns( px, kn=p.kn ) +
+                    Ns( dur, kn=d.kn ):Ns( px, kn=p.kn ),
+            offset = log(Y/10^5),
+            family = poisson,
+              data = Dtb )
> round( ci.exp( mdp ), 3 )
```

```
                                       exp(Est.)   2.5%  97.5%
(Intercept)                                5.617  5.288  5.966
Ns(ax, kn = a.kn)1                         1.080  0.986  1.184
Ns(ax, kn = a.kn)2                         2.080  1.914  2.261
Ns(ax, kn = a.kn)3                         1.066  0.992  1.145
sexF                                       0.718  0.683  0.755
RegionAfrica                              90.915 84.859 97.405
RegionAsia                                15.832 14.608 17.159
RegionOther                                6.120  5.632  6.651
stateDM                                    3.712  2.669  5.162
RegionAfrica:stateDM                       0.306  0.184  0.509
RegionAsia:stateDM                         0.723  0.486  1.074
```

```
RegionOther:stateDM                        1.163  0.765  1.768
stateWell:Ns(px, kn = p.kn)1               0.585  0.532  0.644
stateDM:Ns(px, kn = p.kn)1                 0.439  0.147  1.311
stateWell:Ns(px, kn = p.kn)2               0.652  0.620  0.685
stateDM:Ns(px, kn = p.kn)2                 0.322  0.153  0.681
Ns(px, kn = p.kn)1:Ns(dur, kn = d.kn)1     0.720  0.197  2.623
Ns(px, kn = p.kn)2:Ns(dur, kn = d.kn)1     0.920  0.488  1.736
Ns(px, kn = p.kn)1:Ns(dur, kn = d.kn)2     0.077  0.008  0.717
Ns(px, kn = p.kn)2:Ns(dur, kn = d.kn)2     2.339  0.484 11.311
Ns(px, kn = p.kn)1:Ns(dur, kn = d.kn)3     0.653  0.248  1.719
Ns(px, kn = p.kn)2:Ns(dur, kn = d.kn)3     0.901  0.553  1.467
```

```
> ci.exp( mdp, subset=c("DM","dur","Well:Ns") )
```

```
                                        exp(Est.)        2.5%      97.5%
stateDM                                3.71229550 2.669473650  5.1624926
RegionAfrica:stateDM                   0.30624560 0.184400183  0.5086023
RegionAsia:stateDM                     0.72250140 0.486104520  1.0738601
RegionOther:stateDM                    1.16289674 0.765095868  1.7675286
stateDM:Ns(px, kn = p.kn)1             0.43887289 0.146919494  1.3109861
stateDM:Ns(px, kn = p.kn)2             0.32245455 0.152625007  0.6812575
Ns(px, kn = p.kn)1:Ns(dur, kn = d.kn)1 0.71968605 0.197451715  2.6231629
Ns(px, kn = p.kn)2:Ns(dur, kn = d.kn)1 0.92006199 0.487537899  1.7363041
Ns(px, kn = p.kn)1:Ns(dur, kn = d.kn)2 0.07664781 0.008189894  0.7173337
Ns(px, kn = p.kn)2:Ns(dur, kn = d.kn)2 2.33851904 0.483501712 11.3105521
Ns(px, kn = p.kn)1:Ns(dur, kn = d.kn)3 0.65347997 0.248376995  1.7193061
Ns(px, kn = p.kn)2:Ns(dur, kn = d.kn)3 0.90069981 0.552948925  1.4671521
stateWell:Ns(px, kn = p.kn)1           0.58546902 0.531998130  0.6443142
stateWell:Ns(px, kn = p.kn)2           0.65170444 0.620075073  0.6849472
```

```
> 1-pchisq(abs(md.dev-mdp$deviance),
+          abs(md.df -mdp$df.res) )
```

```
[1] 0
```

We see there is a significant interaction, but we want to show the estimated duration effects, that is the TB RR between DM and non-DM persons as a function of diabetes duration, for different calendar times. However, this is the usual double time scale problem all over again: both diabetes duration and calendar time advance at the same pace. Hence, the effects are better shown as a curve for a set of fixed date of diagnosis of DM, the curves being the joint effect of duration and calendar time.

Therefore we first make an empty plot and then make a loop over dates of DM diagnosis and for each of these compute the RR as a function of duration of diabetes, but only for the period where we actually have observations:

```
> tmpl <- function(ci=FALSE){
+ par( mar=c(3,3,1,1), mgp=c(3,1,0)/1.6 )
+ matplot( 0:16, 0:16, type="n",
+          log="y",las=1, ylim=c(0.5,5),
+          xlab="DM duration (years)", ylab="RR of TB versus non-DM persons")
+ abline( v=seq(0,16,2), h=c(1:15/10,2:10), col=gray(0.8) )
+ abline( h= 1 )
+ box()
+ round( ci.exp( mdp ), 3 )
+ round( ci.exp( mdp, subset=c("DM","dur") ), 3 )
+ for( yod in 1995:2009 )
+ {
+ d.pt <- seq(  0,2010-yod,0.1)
+ p.pt <- seq(yod,2010    ,0.1)
+ Cp <- Ns( p.pt, kn=p.kn)
+ CM <- model.matrix( ~ Ns( p.pt, kn=p.kn ):Ns( d.pt, kn=d.kn) )[,-1]
+ d.eff <- ci.exp( mdp, subset=c("DM","dur"), ctr.mat=cbind(1,0,0,0,Cp,CM) )
+ matlines( d.pt, d.eff,
+           type="l", lty=if(!ci) c(1,0,0) else 1, lwd=c(3,1,1),
+            col=gray(1-(2015-yod)/22) )
+ np <- length( d.pt )
+ if( (yod %% 2) == 1 )
+ text( d.pt[np], d.eff[np,1], paste(yod), adj=c(0,1), font=2,
+        col=gray(1-(2015-yod)/22) )
+ } }
> tmpl()
```

```
> tmpl(ci=TRUE)
```

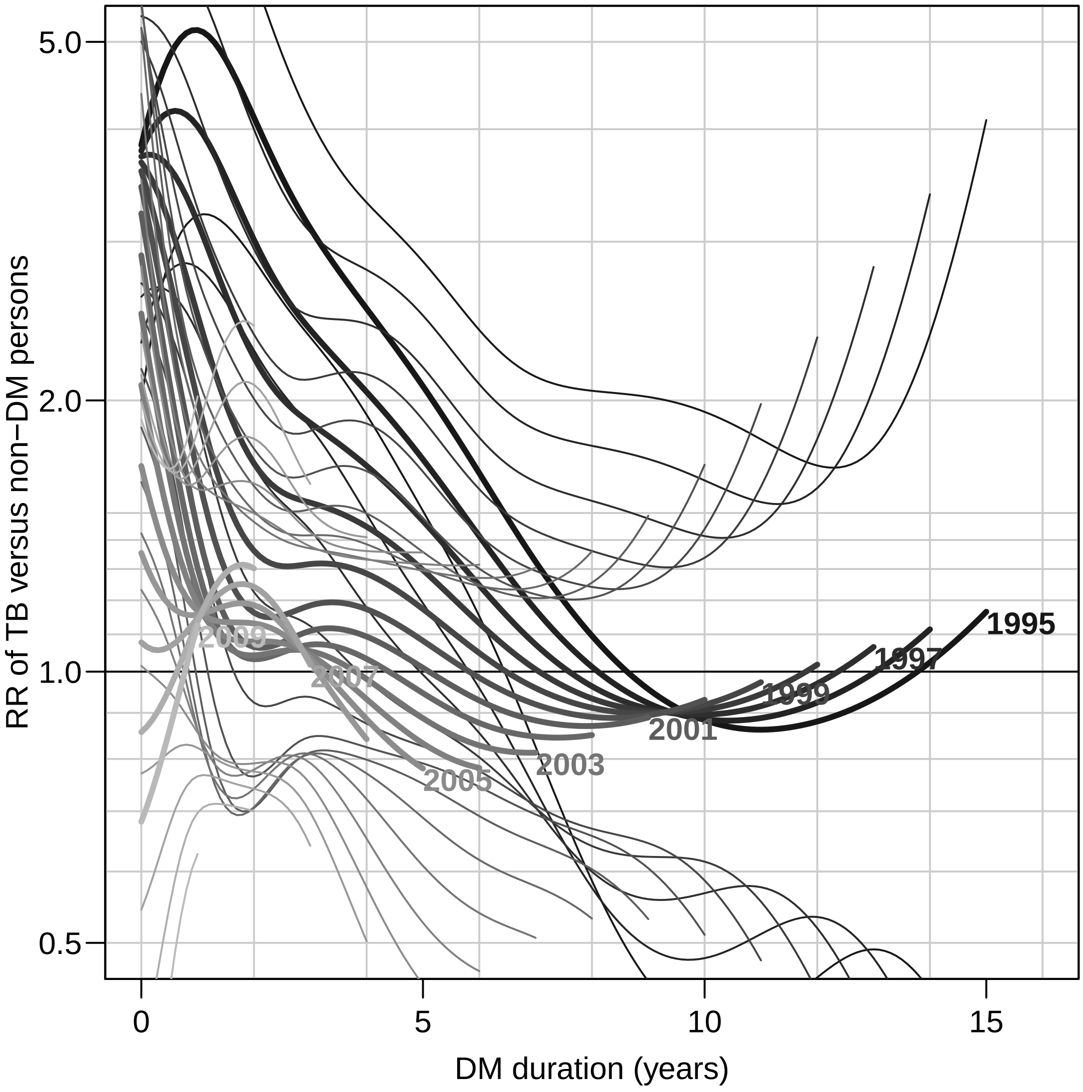

Figure 4.3: *TB-rate-ratios between persons with and without diabetes as a function of diabetes duration. Thin lines are 95% confidence intervals.*

```
> pdf( "../graph/Fig5.pdf", width=8, height=8, pointsize=14 )
> tmpl()
> dev.off()
> postscript( "../graph/Fig5.eps", width=8, height=8, pointsize=14 )
> tmpl()
> dev.off()
```

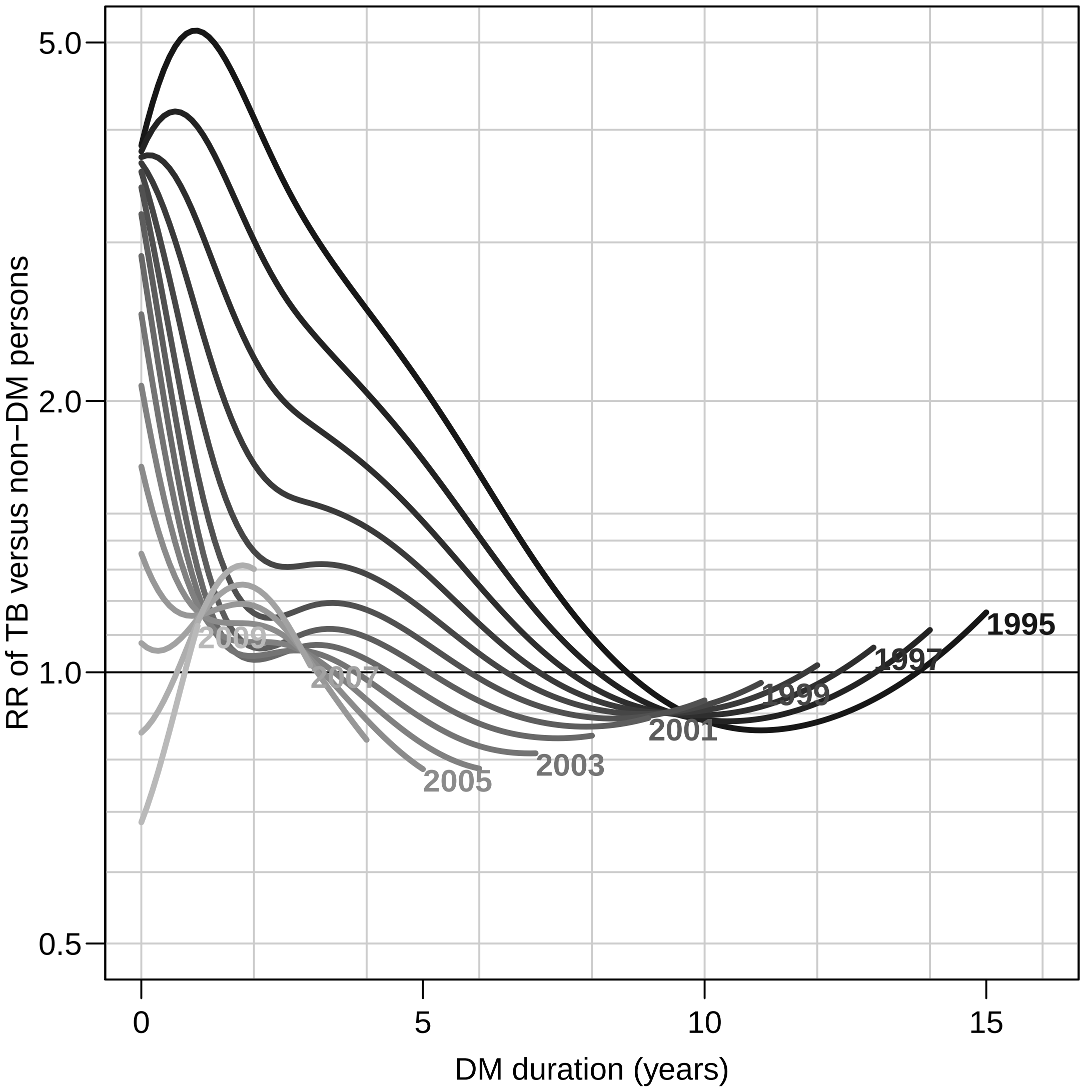

Figure 4.4: *TB-rate-ratios between persons with and without diabetes as a function of diabetes duration.*

```
> win.metafile( "../graph/Fig5.emf", width=8, height=8, pointsize=14 )
> tmpl()
> dev.off()
```

Here is a version of the code that produces a “film” with the effect for each year of diagnosis seprately on a page of a pdf-file:

```
> pdf( "../graph/dmtb-int-film.pdf", width=11, height=11/sqrt(2) )
> for( yod in seq(1995,2009,0.2) )
+ {
+ par( mar=c(3,3,1,1), mgp=c(3,1,0)/1.6 )
+ matplot( 0:16, 0:16, type="n",
+          log="y",las=1, ylim=c(0.4,7),
+          xlab="DM duration (years)", ylab="RR of TB versus non-DM persons")
+ abline( v=seq(0,16,2), h=c(1:15/10,2:10), col=gray(0.8) )
+ abline( h= 1 )
+ box()
+ round( ci.exp( mdp ), 3 )
+ round( ci.exp( mdp, subset=c("DM","dur") ), 3 )
+ d.pt <- seq(  0,2010-yod,0.1)
+ p.pt <- seq(yod,2010    ,0.1)
+ Cp <- Ns( p.pt, kn=p.kn)
+ CM <- model.matrix( ~ Ns( p.pt, kn=p.kn ):Ns( d.pt, kn=d.kn) )[,-1]
+ d.ef95 <- ci.exp( mdp, subset=c("DM","dur"), ctr.mat=cbind(1,0,0,0,Cp,CM) )
+ d.ef90 <- ci.exp( mdp, subset=c("DM","dur"),
+                  ctr.mat=cbind(1,0,0,0,Cp,CM), alpha=0.1 )
+ polygon( c(d.pt,rev(d.pt)), c(d.ef95[,2],rev(d.ef95[,3])),
+          col=gray(0.8), border="transparent" )
+ polygon( c(d.pt,rev(d.pt)), c(d.ef90[,2],rev(d.ef90[,3])),
+          col=gray(0.7), border="transparent" )
+ lines( d.pt, d.ef90[,1], lwd=3, lty=1 )
+ np <- length( d.pt )
+ text( d.pt[np], d.ef90[np,1], paste(yod), adj=c(0,1), font=2 )
+ }
> dev.off()
null device
          1
```

From figure 4.4 it is clear that the calendar time effects mainly has pushed the relative TB incidence down for the longer durations of DM, but also that the RR associated with diabetes for the latest few years (after 2005) have sunk below 2 in the first year after diagnosis of diabetes.

This characteristic patterns of RR with decline immediately after diagnosis of DM is seen for other diseases too, and is most likely a diagnostic artifact and not a biological effect.