Clinical nephropathy from SDC

SDC
July 2016
http://bendixcarstensen.com/
Version 5

Compiled Tuesday 19th July, 2016, 11:26
from: /home/bendix/sdc/proj/HKPWH/SDC.tex

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

Dorte Vistisen — Steno Diabetes Center, Gentofte, Denmark
dtvs@steno.dk

Gregers Stig Andersen — Steno Diabetes Center, Gentofte, Denmark
gsa@steno.dk

Contents

Chapter 1

Reading SDC data

1.1 Reading the SDC clinical data

We have gathered data from the EPR system at SDC — clinical measurements and status of all patients in the EPR system and records of deaths and occurrences of ESRD (dialysis, kidney transplant) derived from the National Patient Register.

1.1.0.1 Utilities

For variable selection and -screening we define a convenience function that prints selected variable names and returns the position of these in the dataframe as a vector — `pat` is an argument in the form of a regular expression:

```
> grnam <- function( pat, dfr, verbose=TRUE )
+ {
+ wh <- grep( pat, names(dfr) )
+ if( verbose ) print( names(dfr)[wh] )
+ return(wh)
+ }
```

. . . and a function that returns the label of the entry from a table of a variable among those with a non-blank label, designed to fish out the most frequently occurring unit name from the lab database:

```
> maxlab <- function( x )
+     {
+     tt <- table(x)
+     tt <- tt[names(tt)!=""]
+     names( tt )[tt==max(tt)]
+     }
```

1.1.1 The `Stata` dataset

We can read the complete datset provided in Stata format, and check that each type of variable actually are in the same type of units:

```
> library( readstata13 )
> library( Epi )
> nef <- read.dta13( "./data/nephrohkworkdata.dta", nonint.factors=TRUE )
> wh <- grnam( "enhed", nef )
```

```
 [1] "abdominalomfang_enhed"      "b12_enhed"
 [3] "blodglukose_enhed"          "bmi_enhed"
 [5] "cpeptid_enhed"              "diastoliskepj_enhed"
 [7] "diurese_enhed"              "dualb_enhed"
 [9] "egfr_enhed"                 "gad_enhed"
[11] "gfr_enhed"                  "haemoglobin_enhed"
[13] "hba1c_enhed"                "hdl_enhed"
[15] "height_enhed"               "hvilepuls_enhed"
[17] "ldl_enhed"                  "middelblodglukoseepj_enhed"
[19] "pcreatinin_enhed"           "systoliskepj_enhed"
[21] "trans_enhed"                "triglycerid_enhed"
[23] "tsh_enhed"                  "ualbcrea_enhed"
[25] "vldl_enhed"                 "weight_enhed"
```

We then list the table for those variables that hev more than one non-blank value, in order to check if any of the variables are recorded in different units:

```
> for( i in wh ) {
+    tt <- table( nef[,i], exclude=NULL )
+    if( length(tt)>3 ){
+      cat( "\n",names(nef)[i],":" )
+      print( tt ) }
+               }
 bmi_enhed :

                         kg/m^2 kg/m<sup>2</sup>             <NA>
          470844          23159             6424                0

 hvilepuls_enhed :

           slag/min slag/min.      <NA>
   497843       548      2036         0

 tsh_enhed :

                    \xd7 10<sup>-3</sup>               miu/l               mlu/l
             432572                    2               51223               16630
               <NA>
                  0
```

and after seeing that all variables are only recorde in one type of unit, we collect these in the object `units`, and remove the corresponding variables from the data frame:

```
> units <- sapply( nef[wh], maxlab )
> names( units ) <- gsub("_enhed","",names(units) )
> cbind( units )
                     units
abdominalomfang      "cm"
b12                  "pmol/l"
blodglukose          "mmol/l"
bmi                  "kg/m^2"
cpeptid              "pmol/l"
diastoliskepj        "mm hg"
diurese              "ml"
dualb                "mg/d"
egfr                 "ml/min"
gad                  "kiu/l"
gfr                  "ml/min"
haemoglobin          "mmol/l"
hba1c                "mmol/mol"
hdl                  "mmol/l"
```

```
height                "m"
hvilepuls             "slag/min."
ldl                   "mmol/l"
middelblodglukoseepj "mmol/l"
pcreatinin            "\xb5mol/l"
systoliskepj          "mm hg"
trans                 "\xb5mol/l"
triglycerid           "mmol/l"
tsh                   "miu/l"
ualbcrea              "mg/g"
vldl                  "mmol/l"
weight                "kg"
> nef <- nef[,-wh]
```

and finally check that we have units of actual variables in `nef`:

```
> match( names(units), names(nef) )
 [1] 16 18 20 21 23 24 25 26 28 29 30 31 32 33 34 35 36 37 40 43 44 45 46 47 48 49
```

Thus we have verified that there are no variables recorded with units differing across the data frame; this is why we could dispense with these variables.

1.2 Dates and events

We produce an overview of the events and -dates, first by listing all variables with a name starting with "d" (this is what the regular expression "^d" means):

```
> wh <- grnam( "^d", nef )
 [1] "dmtype"         "dob"            "debut_diabetes" "date"           "d_esrd"
 [6] "d_renaldisease" "dth"            "d_dth"          "d_stenostart"   "d_stenoslut"
[11] "diastoliskepj"  "diurese"        "dualb"          "duplicates"
> wh <- wh[c(2:6,8:10)]
```

We want more intuitive date names, so we rename the date variables (and `renaldisease` to `ckd` (chronic kidney disease)):

```
> old <- c("dob","debut_diabetes","d_stenostart","date",
+          "d_renaldisease","d_esrd","d_stenoslut","d_dth",
+          "renaldisease")
> new <- c("dob","doDM","doin","dolab",
+          "dockd","doesrd","dox","dodth",
+          "ckd")
> wh <- match( old, names(nef) )
> cbind( names( nef )[wh],   new )
                      new
 [1,] "dob"            "dob"
 [2,] "debut_diabetes" "doDM"
 [3,] "d_stenostart"   "doin"
 [4,] "date"           "dolab"
 [5,] "d_renaldisease" "dockd"
 [6,] "d_esrd"         "doesrd"
 [7,] "d_stenoslut"    "dox"
 [8,] "d_dth"          "dodth"
 [9,] "renaldisease"   "ckd"
>        names( nef )[wh] <- new
```

For further simplification of date handling we transform all date variables to `cal.yr` format. For the variable `doDM` which is merely a numerical variable, we make a copy `dodm` which we make of class `cal.yr`. Thus we preserve the old (partly missing) version in the numerical variable `doDM`):

```
> nef <- cal.yr( nef )
> nef$dodm <- nef$doDM
> class( nef$dodm ) <- class( nef$doDM ) <- class( nef$dob )
```

1.2.1 Date problems

Some of the dates should be known for all, but seem not to be:

```
> wh <- grnam( "^do", nef )
[1] "dob"    "doDM"   "dolab"  "doesrd" "dockd"  "dodth"  "doin"   "dox"    "dodm"
> summary( nef[,wh] )
      dob            doDM          dolab          doesrd          dockd
 Min.   :1901   Min.   :1933   Min.   :1913   Min.   :1979   Min.   :1979
 1st Qu.:1940   1st Qu.:1979   1st Qu.:2002   1st Qu.:2004   1st Qu.:2005
 Median :1950   Median :1990   Median :2007   Median :2009   Median :2008
 Mean   :1952   Mean   :1987   Mean   :2007   Mean   :2008   Mean   :2008
 3rd Qu.:1964   3rd Qu.:1998   3rd Qu.:2011   3rd Qu.:2013   3rd Qu.:2012
 Max.   :2000   Max.   :2014   Max.   :2015   Max.   :2015   Max.   :2015
                NA's   :15984                 NA's   :495427 NA's   :464292
     dodth            doin           dox            dodm
 Min.   :2001   Min.   :1988   Min.   :1994   Min.   :1933
 1st Qu.:2005   1st Qu.:1994   1st Qu.:2007   1st Qu.:1979
 Median :2009   Median :1998   Median :2010   Median :1990
 Mean   :2009   Mean   :2000   Mean   :2010   Mean   :1987
 3rd Qu.:2012   3rd Qu.:2004   3rd Qu.:2013   3rd Qu.:1998
 Max.   :2015   Max.   :2015   Max.   :2015   Max.   :2014
 NA's   :497078 NA's   :1524   NA's   :272129 NA's   :15984
```

There is clearly a wrongly coded date in `dolab`, which we remove:

```
> subset( nef, dolab < 1930 )
       newid  sex dmtype      dob doDM    dolab       age esrd doesrd ckd dockd dth dodth
147550  4516 Male type 2 1946.082 1998 1913.217 -32.86516    0     NA   0    NA   0    NA
           doin dox abdominalomfang           alkohol b12 black blodglukose bmi
147550 2011.925  NA              NA <14 Genstande/uge  NA     0          NA  NA
       civilstandskode cpeptid diastoliskepj diurese dualb duplicates egfr gad gfr
147550                      NA            NA      NA    NA          0   NA  NA  NA
       haemoglobin hba1c hdl height hvilepuls ldl middelblodglukoseepj migrant
147550          NA    NA  NA     NA        NA  NA                   NA       0
                motion pcreatinin  region rygning systoliskepj trans triglycerid tsh
147550 Genoptr\xe6ning         NA Denmark                   NA    NA          NA  NA
       ualbcrea vldl weight dodm
147550       NA   NA     NA 1998
> nef <- subset( nef, dolab > 1930)
```

There are missing values for date of diabetes (`doDM`) and also date of entry to SDC, `doin`.

```
> tt <- with( nef, table(newid,is.na(doDM)) )
> dim(tt)
[1] 15210     2
> range( apply( tt>0, 1, sum ) )
```

```
[1] 1 1
> apply( tt>0, 2, sum )
FALSE  TRUE
12955  2255
```

Thus we see that there no persons with both missing and non-missing values of `doDM` in their records, and hat there are 2255 persons with missing date of DM, and hence unknown diabetes duration for something in the vicinity of 20% of the persons in the data set.

We make a check for the other date variables with missing values, to see if missing and non-missing values occur within the same person. To this end we devise a function that first computes the number of missing and non-missing values for each variable and persons, and then how many persons have both missing and non-missing values for each of the variables:

```
> na.chk <- function( var )
+     {
+     tt <- table( nef$newid, is.na(nef[,var] ) )
+     print( sum( apply( tt>0, 1, sum ) > 1 ) )
+     invisible( tt )
+     }
> t.ren <- na.chk("dockd")
[1] 4229
> t.esr <- na.chk("doesrd")
[1] 477
> t.dth <- na.chk("dodth")
[1] 3332
> t.dm <- na.chk("doDM")
[1] 0
> t.in <- na.chk("doin")
[1] 0
> t.ex <- na.chk("dox")
[1] 0
```

We see that `doDM`, `doin` and `dox` are either missing non-missing for all records from the same person. Moreover, the non-missing values are identical within persons:

```
> summary( with( nef, tapply( doDM, newid, var, na.rm=TRUE ) ) )
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
      0       0       0       0       0       0    2288
> summary( with( nef, tapply( doin, newid, var, na.rm=TRUE ) ) )
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
      0       0       0       0       0       0     576
> summary( with( nef, tapply( dox , newid, var, na.rm=TRUE ) ) )
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
      0       0       0       0       0       0    6222
```

But this is not the case with the other date variables (`dockd`, `doesrd` and `dodth`); person no. 44 illustrates what the real structure of the data is for these:

```
> wh <- c("newid","dob","doDM","dolab","ckd","dockd","esrd","doesrd","dth","dodth")
> head( nef[nef$newid==44,wh] )
```

```
     newid      dob doDM    dolab ckd    dockd esrd doesrd dth dodth
1288    44 1966.619 1996 1998.222   1 1998.222    0     NA   0    NA
1289    44 1966.619 1996 1998.512   1 1998.512    0     NA   0    NA
1290    44 1966.619 1996 1998.778   0       NA    0     NA   0    NA
1291    44 1966.619 1996 1998.780   1 1998.780    0     NA   0    NA
1292    44 1966.619 1996 1998.882   1 1998.882    0     NA   0    NA
1293    44 1966.619 1996 1999.142   0       NA    0     NA   0    NA
> tail( nef[nef$newid==44,wh] )
     newid      dob doDM    dolab ckd    dockd esrd   doesrd dth    dodth
1330    44 1966.619 1996 2002.014   1 2002.014    1 2002.014   0       NA
1331    44 1966.619 1996 2002.041   1 2002.041    1 2002.041   0       NA
1332    44 1966.619 1996 2002.115   1 2002.115    1 2002.115   0       NA
1333    44 1966.619 1996 2002.120   1 2002.120    1 2002.120   0       NA
1334    44 1966.619 1996 2002.227   1 2002.227    1 2002.227   0       NA
1335    44 1966.619 1996 2002.238   0       NA    0       NA   1 2002.238
```

The non-missing values of the dates `dockd`, `doesrd` and `dodth` are always identical to `dolab`, so what we really need is to change these to the earliest date for each person. This means that there will then be two possible indicators of ESRD (and similarly CKD) available, namely:

- `esrd` indicating whether a person meet the criteria for ESRD *at* the date of visit (`dolab`)
- the logical (`dolab`≥`doesrd`) indicating whether a person has met the ESRD criteria at least once prior to the current visit date.

In order to obtain this we use the `ave` function — and also a version of the min function that ignores NAs and for an all-NA input returns NA (instead of Inf, which logically *is* the minimum of the NULL object left after removing the NAs):

```
> miNA <- function(x) if( all(is.na(x)) ) NA else min( x, na.rm=TRUE )
> for( vv in c("doesrd","dockd","dodth") )
+    nef[,vv] <- ave( nef[,vv], nef$newid, FUN = miNA )
> head( nef[nef$newid==44,wh] )
     newid      dob doDM    dolab ckd    dockd esrd   doesrd dth    dodth
1288    44 1966.619 1996 1998.222   1 1998.222    0 1999.667   0 2002.238
1289    44 1966.619 1996 1998.512   1 1998.222    0 1999.667   0 2002.238
1290    44 1966.619 1996 1998.778   0 1998.222    0 1999.667   0 2002.238
1291    44 1966.619 1996 1998.780   1 1998.222    0 1999.667   0 2002.238
1292    44 1966.619 1996 1998.882   1 1998.222    0 1999.667   0 2002.238
1293    44 1966.619 1996 1999.142   0 1998.222    0 1999.667   0 2002.238
> tail( nef[nef$newid==44,wh] )
     newid      dob doDM    dolab ckd    dockd esrd   doesrd dth    dodth
1330    44 1966.619 1996 2002.014   1 1998.222    1 1999.667   0 2002.238
1331    44 1966.619 1996 2002.041   1 1998.222    1 1999.667   0 2002.238
1332    44 1966.619 1996 2002.115   1 1998.222    1 1999.667   0 2002.238
1333    44 1966.619 1996 2002.120   1 1998.222    1 1999.667   0 2002.238
1334    44 1966.619 1996 2002.227   1 1998.222    1 1999.667   0 2002.238
1335    44 1966.619 1996 2002.238   0 1998.222    0 1999.667   1 2002.238
```

In order to remedy the missing dates of DM, we impute as date of diabetes 3 months before the first known visit, and also backdating those values of date of diagnosis that are *later* than the earliest known visit:

```
> nef$dodm <- ifelse( is.na(nef$doDM) |
+                           nef$doDM > ave( nef$dolab, nef$newid, FUN=min ),
+                     ave( nef$dolab, nef$newid, FUN=min ) - 1/4,
+                     nef$doDM )
```

After this, `doDM` is the orginal incomplete (and partly non-credible) date of diagnosis, and `dodm` the revised version that is guaranteed to be before the first recorded visit.

We also exclude visits prior to 2001, since we do not have any deaths recorded before 2001 — the earliest is `min(as.Date.cal.yr(nef$dodth),na.rm=TRUE)`= 2001-03-29. Thus any mesurements before 2001 (we will use 1 January 2001 as cutpoint) will be among people that are known to be alive in 2001, and therefore likely biased. This consitutes a fair chunk:

```
> nrow( nef )
[1] 500426
> nef <- subset( nef, dolab>2001 )
> nrow( nef )
[1] 417462
```

After this exercise the dates should ideally be in the following order:

$$\texttt{dobth} < \texttt{dodm} < \texttt{doin} < \texttt{dolab} < \texttt{dox} \leq \texttt{dodth}$$

and for the disease outcomes:

$$\texttt{dodm} < \texttt{dockd} \leq \texttt{doesrd} < \texttt{dodth}$$

Now, only the `dolab` varies between visits, all the other dates are identical within persons.

We should not have any records with a valid date of event equal to visit data and 0 in event indicator; but apparently this does occur:

```
> with( nef, cbind(
+             table( dolab==dockd , ckd , exclude=NULL ),
+             table( dolab==doesrd, esrd, exclude=NULL ),
+             table( dolab==dodth , dth , exclude=NULL ) ) )
            0     1 <NA>      0    1 <NA>      0    1 <NA>
FALSE  124682 30010    0  14775 4040    0  73263    0    0
TRUE        0  3429    0      3  399    0      7 3349    0
<NA>   259341     0    0 398245    0    0 340843    0    0
```

```
> ( zz <- subset( nef, (dolab==dockd  &  ckd==0) |
+                      (dolab==doesrd & esrd==0) |
+                      (dolab==dodth  &  dth==0) )[,wh] )
       newid      dob doDM    dolab ckd    dockd esrd   doesrd dth    dodth
16887    548 1946.118 1992 2013.102   0       NA    0       NA   0 2013.102
72866   2248 1971.153 1981 2002.482   1 2001.770    0 2002.482   0       NA
104323  3219 1927.730 1999 2012.185   0       NA    0       NA   0 2012.185
113689  3515 1928.601 1992 2005.091   1 1998.550    1 2002.444   0 2005.091
119719  3690 1926.068 1968 2012.798   0 2004.604    0       NA   0 2012.798
183520  5616 1927.747 1974 2002.249   0       NA    0       NA   0 2002.249
233690  7123 1937.118 1990 2002.687   0 1999.927    0       NA   0 2002.687
243772  7448 1975.388 1984 2004.858   1 2004.858    0 2004.858   0 2005.926
263668  8072 1927.703   NA 2001.173   1 1998.438    0 2001.173   0 2001.439
429411 13147 1968.094 1973 2011.136   1 1999.873    1 1999.873   0 2011.136
```

```
> fishy <- subset( nef, ( dolab==doesrd | dolab==dodth ) & newid %in% zz$newid )
> for( ii in zz$newid ) print( subset(fishy,newid==ii) )
      newid    sex dmtype      dob doDM    dolab      age esrd doesrd ckd dockd dth
16886   548 Female type 2 1946.118 1992 2013.102 66.98426    0     NA   0    NA   1
16887   548 Female type 2 1946.118 1992 2013.102 66.98426    0     NA   0    NA   0
        dodth     doin      dox abdominalomfang alkohol b12 black blodglukose bmi
16886 2013.102 1998.214 2013.102              NA          NA     0          NA  NA
16887 2013.102 1998.214 2013.102              NA          NA     0          NA  NA
      civilstandskode cpeptid diastoliskepj diurese dualb duplicates egfr gad gfr
16886                      NA            NA      NA    NA          0   NA  NA  NA
16887                      NA            NA      NA    NA          0   NA  NA  NA
      haemoglobin hba1c hdl height hvilepuls ldl middelblodglukoseepj migrant motion
16886          NA    NA  NA     NA        NA  NA                   NA       0
16887          NA    NA  NA     NA        NA  NA                   NA       0
      pcreatinin  region    rygning systoliskepj trans triglycerid tsh ualbcrea vldl
16886         NA Denmark                     NA    NA          NA  NA       NA   NA
16887         NA Denmark Ikke ryger          NA    NA          NA  NA       NA   NA
      weight dodm
16886     NA 1992
16887     NA 1992
      newid  sex dmtype      dob doDM    dolab      age esrd   doesrd ckd   dockd dth
72865  2248 Male type 1 1971.153 1981 2002.482 31.32923    1 2002.482   1 2001.77   0
72866  2248 Male type 1 1971.153 1981 2002.482 31.32923    0 2002.482   1 2001.77   0
      dodth     doin      dox abdominalomfang alkohol b12 black blodglukose bmi
72865    NA 2001.732 2012.757              NA          NA    NA          NA  NA
72866    NA 2001.732 2012.757              NA          NA    NA          NA  NA
      civilstandskode cpeptid diastoliskepj diurese dualb duplicates egfr gad gfr
72865                      NA            NA      NA    NA          0   NA  NA  NA
72866                      NA            NA      NA    NA          0   NA  NA  NA
      haemoglobin hba1c hdl height hvilepuls ldl middelblodglukoseepj migrant motion
72865          NA    NA  NA     NA        NA  NA                   NA      NA
72866          NA    NA  NA     NA        NA  NA                   NA      NA
      pcreatinin region rygning systoliskepj trans triglycerid tsh ualbcrea vldl weight
72865         NA   <NA>                   NA    NA          NA  NA       NA   NA     NA
72866         NA   <NA>                   NA    NA          NA  NA       NA   NA     NA
      dodm
72865 1981
72866 1981
       newid    sex dmtype     dob doDM    dolab      age esrd doesrd ckd dockd dth
104323  3219 Female type 2 1927.73 1999 2012.185 84.45448    0     NA   0    NA   0
104324  3219 Female type 2 1927.73 1999 2012.185 84.45448    0     NA   0    NA   1
         dodth     doin      dox abdominalomfang alkohol b12 black blodglukose bmi
104323 2012.185 2004.204 2004.875              NA          NA     0          NA  NA
104324 2012.185 2004.204 2004.875              NA          NA     0          NA  NA
       civilstandskode cpeptid diastoliskepj diurese dualb duplicates egfr gad gfr
104323                     NA            NA      NA    NA          0   NA  NA  NA
104324                     NA            NA      NA    NA          0   NA  NA  NA
       haemoglobin hba1c hdl height hvilepuls ldl middelblodglukoseepj migrant motion
104323          NA    NA  NA     NA        NA  NA                   NA       0
104324          NA    NA  NA     NA        NA  NA                   NA       0
       pcreatinin  region    rygning systoliskepj trans triglycerid tsh ualbcrea vldl
104323         NA Denmark Ikke ryger          NA    NA          NA  NA       NA   NA
104324         NA Denmark                     NA    NA          NA  NA       NA   NA
       weight dodm
104323     NA 1999
104324     NA 1999
       newid    sex dmtype      dob doDM    dolab      age esrd   doesrd ckd   dockd dth
113662  3515 Female type 2 1928.601 1992 2002.444 73.84258    1 2002.444   1 1998.55   0
113688  3515 Female type 2 1928.601 1992 2005.091 76.49007    0 2002.444   0 1998.55   1
```

```
113689  3515 Female type 2 1928.601 1992 2005.091 76.49007    1 2002.444   1 1998.55   0
          dodth     doin      dox abdominalomfang alkohol b12 black blodglukose bmi
113662 2005.091 2001.269 2005.091              NA          NA     0          NA  31
113688 2005.091 2001.269 2005.091              NA          NA     0          NA  NA
113689 2005.091 2001.269 2005.091              NA          NA    NA          NA  NA
       civilstandskode cpeptid diastoliskepj diurese dualb duplicates     egfr gad gfr
113662          D\xf8d    2510            NA      NA    NA          0 12.31769  NA  NA
113688                      NA            NA      NA    NA          0       NA  NA  NA
113689                      NA            NA      NA    NA          0       NA  NA  NA
       haemoglobin hba1c  hdl height hvilepuls ldl middelblodglukoseepj migrant motion
113662         6.4    79 1.38   1.68        NA  NA                    NA       0
113688          NA    NA   NA     NA        NA  NA                    NA       0
113689          NA    NA   NA     NA        NA  NA                    NA      NA
       pcreatinin  region rygning systoliskepj trans triglycerid tsh ualbcrea vldl weight
113662        333 Denmark                   NA    39        5.79 3.1       NA   NA   87.8
113688         NA Denmark                   NA    NA          NA  NA       NA   NA     NA
113689         NA    <NA>                   NA    NA          NA  NA       NA   NA     NA
       dodm
113662 1992
113688 1992
113689 1992
       newid    sex dmtype      dob doDM    dolab      age esrd doesrd ckd    dockd dth
119718  3690 Female type 1 1926.068 1968 2012.798 86.72964    0     NA   0 2004.604   1
119719  3690 Female type 1 1926.068 1968 2012.798 86.72964    0     NA   0 2004.604   0
          dodth     doin      dox abdominalomfang alkohol b12 black blodglukose bmi
119718 2012.798 1993.754 2012.798              NA          NA     0          NA  NA
119719 2012.798 1993.754 2012.798              NA          NA     0          NA  NA
       civilstandskode cpeptid diastoliskepj diurese dualb duplicates egfr gad gfr
119718                      NA            NA      NA    NA          0   NA  NA  NA
119719                      NA            NA      NA    NA          0   NA  NA  NA
       haemoglobin hba1c hdl height hvilepuls ldl middelblodglukoseepj migrant motion
119718          NA    NA  NA     NA        NA  NA                    NA       0
119719          NA    NA  NA     NA        NA  NA                    NA       0
       pcreatinin  region    rygning systoliskepj trans triglycerid tsh ualbcrea vldl
119718         NA Denmark                     NA    NA          NA  NA       NA   NA
119719         NA Denmark Ikke ryger          NA    NA          NA  NA       NA   NA
       weight dodm
119718     NA 1968
119719     NA 1968
       newid  sex            dmtype      dob doDM    dolab     age esrd doesrd ckd dockd
183520  5616 Male type ikke angivet 1927.747 1974 2002.249 74.5024    0     NA   0    NA
183521  5616 Male type ikke angivet 1927.747 1974 2002.249 74.5024    0     NA   0    NA
       dth    dodth     doin      dox abdominalomfang alkohol b12 black blodglukose bmi
183520   0 2002.249 1993.754 2002.249              NA          NA     0          NA  NA
183521   1 2002.249 1993.754 2002.249              NA          NA     0          NA  NA
       civilstandskode cpeptid diastoliskepj diurese dualb duplicates egfr gad gfr
183520                      NA            NA      NA    NA          0   NA  NA  NA
183521                      NA            NA      NA    NA          0   NA  NA  NA
       haemoglobin hba1c hdl height hvilepuls ldl middelblodglukoseepj migrant motion
183520          NA    NA  NA     NA        NA  NA                    NA       0  Ingen
183521          NA    NA  NA     NA        NA  NA                    NA       0
       pcreatinin  region           rygning systoliskepj trans triglycerid tsh ualbcrea
183520         NA Denmark <3 cigaretter/dag           NA    NA          NA  NA       NA
183521         NA Denmark                             NA    NA          NA  NA       NA
       vldl weight dodm
183520   NA     NA 1974
183521   NA     NA 1974
       newid  sex            dmtype      dob doDM    dolab      age esrd doesrd ckd
233690  7123 Male type ikke angivet 1937.118 1990 2002.687 65.56879    0     NA   0
```

```
233691  7123 Male type ikke angivet 1937.118 1990 2002.687 65.56879    0     NA   0
          dockd dth    dodth     doin      dox abdominalomfang alkohol b12 black
233690 1999.927   0 2002.687 1993.754 2002.687              NA          NA     0
233691 1999.927   1 2002.687 1993.754 2002.687              NA          NA     0
       blodglukose bmi civilstandskode cpeptid diastoliskepj diurese dualb duplicates
233690          NA  NA          D\xf8d      NA            NA      NA    NA          0
233691          NA  NA                      NA            NA      NA    NA          0
           egfr gad gfr haemoglobin hba1c hdl height hvilepuls ldl middelblodglukoseepj
233690 88.95801  NA  NA          NA    73  NA     NA        NA  NA                   NA
233691       NA  NA  NA          NA    NA  NA     NA        NA  NA                   NA
       migrant motion pcreatinin  region rygning systoliskepj trans triglycerid tsh
233690       0               NA Denmark                   NA    NA          NA  NA
233691       0               NA Denmark                   NA    NA          NA  NA
       ualbcrea vldl weight dodm
233690       NA   NA     NA 1990
233691       NA   NA     NA 1990
       newid  sex dmtype      dob doDM    dolab      age esrd   doesrd ckd    dockd dth
243771  7448 Male type 1 1975.388 1984 2004.858 29.47023    1 2004.858   1 2004.858   0
243772  7448 Male type 1 1975.388 1984 2004.858 29.47023    0 2004.858   1 2004.858   0
243777  7448 Male type 1 1975.388 1984 2005.926 30.53799    0 2004.858   0 2004.858   1
          dodth     doin     dox abdominalomfang alkohol b12 black blodglukose bmi
243771 2005.926 2001.921 2002.69              NA          NA    NA          NA  NA
243772 2005.926 2001.921 2002.69              NA          NA    NA          NA  NA
243777 2005.926 2001.921 2002.69              NA          NA     0          NA  NA
       civilstandskode cpeptid diastoliskepj diurese dualb duplicates egfr gad gfr
243771                      NA            NA      NA    NA          0   NA  NA  NA
243772                      NA            NA      NA    NA          0   NA  NA  NA
243777                      NA            NA      NA    NA          0   NA  NA  NA
       haemoglobin hba1c hdl height hvilepuls ldl middelblodglukoseepj migrant motion
243771          NA    NA  NA     NA        NA  NA                   NA      NA
243772          NA    NA  NA     NA        NA  NA                   NA      NA
243777          NA    NA  NA     NA        NA  NA                   NA       0
       pcreatinin  region rygning systoliskepj trans triglycerid tsh ualbcrea vldl weight
243771         NA    <NA>                   NA    NA          NA  NA       NA   NA     NA
243772         NA    <NA>                   NA    NA          NA  NA       NA   NA     NA
243777         NA Denmark                    NA    NA          NA  NA       NA   NA     NA
       dodm
243771 1984
243772 1984
243777 1984
       newid  sex dmtype      dob doDM    dolab      age esrd   doesrd ckd    dockd dth
263668  8072 Male type 2 1927.703   NA 2001.173 73.47022    0 2001.173   1 1998.438   0
263669  8072 Male type 2 1927.703   NA 2001.173 73.47022    1 2001.173   1 1998.438   0
263671  8072 Male type 2 1927.703   NA 2001.439 73.73579    0 2001.173   0 1998.438   1
          dodth     doin      dox abdominalomfang alkohol b12 black blodglukose bmi
263668 2001.439 1993.754 1998.742              NA          NA    NA          NA  NA
263669 2001.439 1993.754 1998.742              NA          NA    NA          NA  NA
263671 2001.439 1993.754 1998.742              NA          NA    NA          NA  NA
       civilstandskode cpeptid diastoliskepj diurese dualb duplicates egfr gad gfr
263668                      NA            NA      NA    NA          0   NA  NA  NA
263669                      NA            NA      NA    NA          0   NA  NA  NA
263671                      NA            NA      NA    NA          0   NA  NA  NA
       haemoglobin hba1c hdl height hvilepuls ldl middelblodglukoseepj migrant motion
263668          NA    NA  NA     NA        NA  NA                   NA      NA
263669          NA    NA  NA     NA        NA  NA                   NA      NA
263671          NA    NA  NA     NA        NA  NA                   NA      NA
       pcreatinin region rygning systoliskepj trans triglycerid tsh ualbcrea vldl weight
263668         NA   <NA>                   NA    NA          NA  NA       NA   NA     NA
263669         NA   <NA>                   NA    NA          NA  NA       NA   NA     NA
```

```
263671         NA   <NA>                    NA    NA          NA  NA       NA   NA     NA
          dodm
263668 1997.903
263669 1997.903
263671 1997.903
       newid    sex dmtype      dob doDM    dolab      age esrd   doesrd ckd    dockd dth
429410 13147 Female type 1 1968.094 1973 2011.136 43.04175    0 1999.873   0 1999.873   1
429411 13147 Female type 1 1968.094 1973 2011.136 43.04175    1 1999.873   1 1999.873   0
          dodth     doin      dox abdominalomfang alkohol b12 black blodglukose bmi
429410 2011.136 1995.884 2011.136              NA          NA     0          NA  NA
429411 2011.136 1995.884 2011.136              NA          NA    NA          NA  NA
       civilstandskode cpeptid diastoliskepj diurese dualb duplicates egfr gad gfr
429410                      NA            NA      NA    NA          0   NA  NA  NA
429411                      NA            NA      NA    NA          0   NA  NA  NA
       haemoglobin hba1c hdl height hvilepuls ldl middelblodglukoseepj migrant motion
429410          NA    NA  NA     NA         NA  NA                   NA       0
429411          NA    NA  NA     NA         NA  NA                   NA      NA
       pcreatinin  region rygning systoliskepj trans triglycerid tsh ualbcrea vldl weight
429410         NA Denmark                   NA    NA          NA  NA       NA   NA     NA
429411         NA    <NA>                   NA    NA          NA  NA       NA   NA     NA
       dodm
429410 1973
429411 1973
```

On inspection it is seen that it all boils down to duplicated records — the same `newid` and `dolab`, but with possibly different status:

```
> table( duplicated(nef[,c("newid","dolab")]), ckd  =nef$ckd  )
       ckd
             0      1
  FALSE 384000  33363
  TRUE      23     76
> table( duplicated(nef[,c("newid","dolab")]), esrd =nef$esrd )
       esrd
             0      1
  FALSE 412974   4389
  TRUE      49     50
> table( duplicated(nef[,c("newid","dolab")]), death=nef$dth  )
       death
             0      1
  FALSE 414017   3346
  TRUE      96      3
```

In order to weed out these duplicate records we take the average of the clinical measurements, and a random of the values of the non-numeric variables:

```
> # where is the key (names in kn) (numeric)
> ix <- match( kn <- c("newid","dolab"), names(nef) )
> # names of numeric variables, except key
> wh <- names( which( sapply( nef[,-ix], is.numeric ) ) )
> # names of non-numeric variables, except key
> hw <- setdiff( names(nef), c(wh,kn) )
> # average over non-missing within key
> system.time( an <- aggregate( nef[,wh], nef[,kn],
+                               FUN = mean, na.rm=TRUE ) )
   user  system elapsed
270.242   0.080 270.244
```

```
> # reset the integers for ckd, esrd and death:
> an[,c("ckd","esrd","dth")] <- ( an[,c("ckd","esrd","dth")] > 0 )*1
> # the first value of the non-numerical variables
> af <- nef[!duplicated(nef[,kn]),c(kn,hw)]
> nrow( nef )
[1] 417462
> nrow( an )
[1] 417363
> nrow( af )
[1] 417363
> intersect( names(af), names(an) )
[1] "newid" "dolab"
> nef <- merge( af, an )
> nrow( nef )
[1] 417363
```

Thus we have now a dataset with key (`newid`,`dolab`).

To inspect the relationship between the other dates we shave the dataset down to one record per person:

```
> # only one record per person
> wh <- grnam( "^do", nef )
[1] "dolab"  "dob"    "doDM"   "doesrd" "dockd"  "dodth"  "doin"   "dox"    "dodm"
> np <- nef[!duplicated(nef$newid),wh]
> # diabetes before birth?
> with( np, table( doDM >= dob, exclude=NULL ) )
FALSE  TRUE  <NA>
    4 12934  2247
> subset( np, doDM < dob )
          dolab      dob doDM   doesrd    dockd    dodth     doin      dox dodm
45538  2010.225 1977.981 1977      NaN      NaN      NaN 2010.225      NaN 1977
93870  2013.943 1994.290 1994      NaN      NaN      NaN 2013.943      NaN 1994
148083 2004.738 1970.509 1970      NaN      NaN      NaN 2004.738 2008.650 1970
305652 2002.402 1964.387 1964 2009.433 2005.028 2013.677 2002.400 2013.677 1964
> # renal disease before DM?
> with( np, table( dockd >= doDM, exclude=NULL ) )
FALSE  TRUE  <NA>
   12  4067 11106
> # ESRD before DM?
> with( np, table( doesrd >= doDM, exclude=NULL ) )
FALSE  TRUE  <NA>
   11   444 14730
> # ESRD before renal disease ?
> with( np, table( doesrd >= dockd, exclude=NULL ) )
 TRUE  <NA>
  478 14707
> # Death after any type of event ?
> with( np, table( dodth >= pmax(doDM,dockd,doesrd,na.rm=TRUE) ) )
TRUE
3186
```

There are a few that are obviously diagnosed as infants, so we re-set their date of diabetes to 3 months after birth:

```
> nef <- transform( nef, doDM = ifelse( doDM<dob, dob+1/4, doDM) )
```

Finally, there are a few persons with entry dates that are clearly too early, as the earliest known is 3rd October 1993, which is used for persons prevalent as SDC pateints at thus date, so we reset these dates to this, and create an indicator variable for this:

```
> tt <- table( np$doin )
> tt[tt==max(tt)]
1993.75359342916
            3114
> as.Date.cal.yr( mostin <- as.numeric(names(tt[tt==max(tt)])) )
[1] "1993-10-03"
> sort( np$doin )[1:10]
 [1] 1988.086 1993.721 1993.737 1993.743 1993.754 1993.754 1993.754 1993.754 1993.754
[10] 1993.754
> nef$doin <- pmax( nef$doin, mostin )
> nef$prev <- ( abs( nef$doin - mostin ) < 0.1 )
> summary( nef$doin )
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   1994    1994    2000    2000    2006    2015    1404
```

And so finally we can save the groomed dataframe:

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

1.3 Overview of dates

We now make histograms of the different dates, so we take the dataset and shave it down to one record per person:

```
> load( file="./data/nef.Rda" )
> nuf <- nef[,c("dob",
+               "doDM",
+               "dodm",
+               "doin",
+               "dockd",
+               "doesrd",
+               "dodth",
+               "dox")]
> dim( nef )
[1] 417363     51
> dim( nuf )
[1] 417363      8
> nuf <- nuf[!duplicated(nuf),]
> dim( nuf )
[1] 15184     8
```

```
> hh <-
+ function( x, lab, ... ) hist(x, col="black", main="", xlab=lab, ylab="", ... )
> par( mfrow=c(3,3), mar=c(3,3,1,1), mgp=c(3,1,0)/1.6, las=1 )
> hh( nuf$dob   , "Date of birth"          , breaks=seq(1900,2016,1   ) )
> hh( nuf$doDM  , "Date of diabetes debut", breaks=seq(1930,2016,1   ), ylim=c(0,600) )
> hh( nuf$dodm  , "Amended diabetes debut", breaks=seq(1930,2016,1   ), ylim=c(0,600) )
> hh( nef$dolab , "Date of visit to SDC"  , breaks=seq(2000,2016,1/12) ) ; abline(v=2000:20
> hh( nuf$dockd , "Date of CKD"           , breaks=seq(1979,2016,1/ 2) ) ; axis(side=1,at=1
> hh( nuf$doesrd, "Date of ESRD"          , breaks=seq(1979,2016,1/ 2) ) ; axis(side=1,at=1
> hh( nuf$dodth , "Date of death"         , breaks=seq(2000,2016,1/12) ) ; abline(v=2000:20
> hh( nuf$doin  , "Date of entry at SDC"  , breaks=seq(1993,2016,1/12), ylim=c(0,200) ) ; a
> hh( nuf$dox   , "Date of exit from SDC"  , breaks=seq(1993,2016,1/12), ylim=c(0,200) ) ; a
```

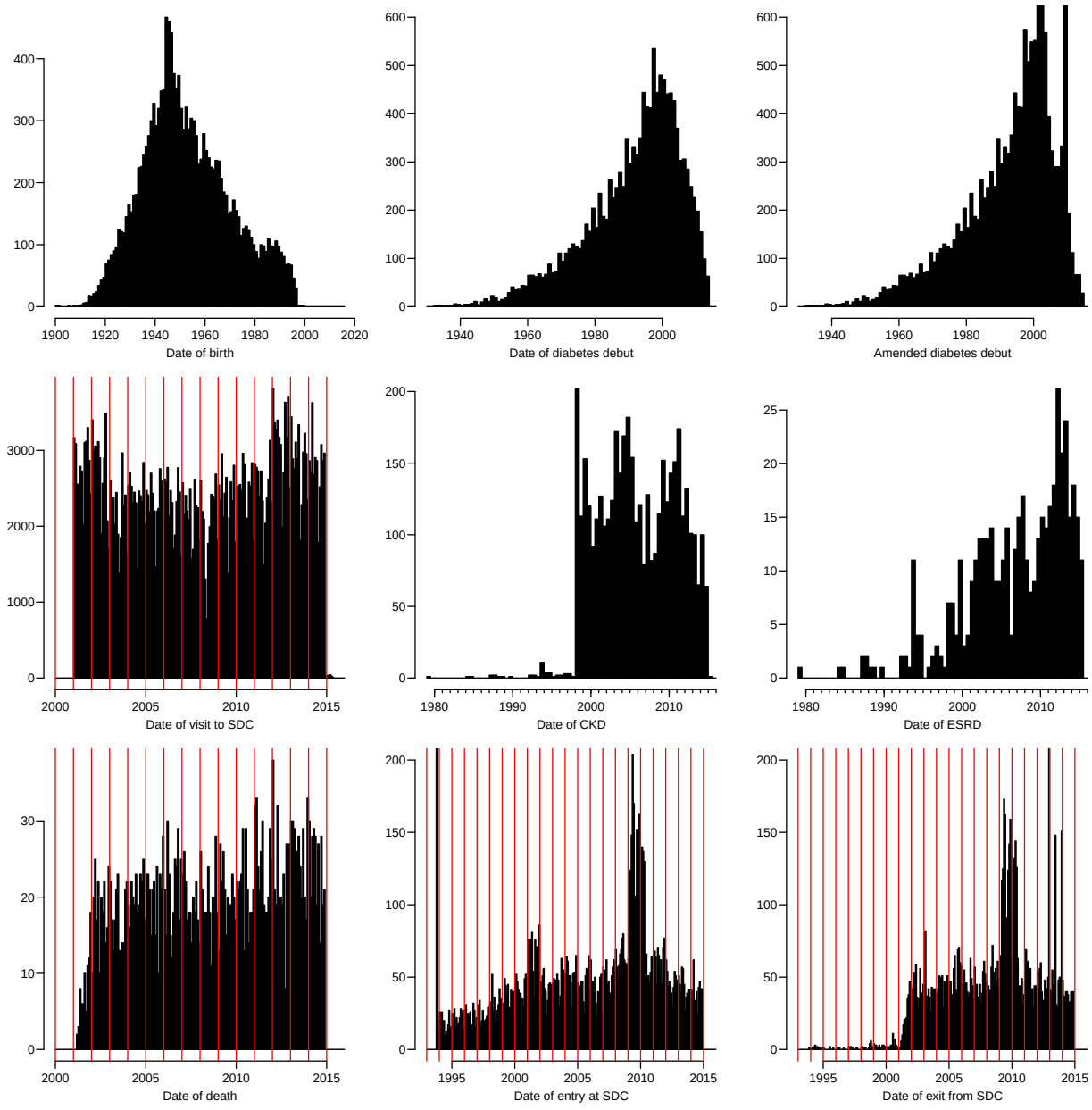

Figure 1.1: *Histograms of various dates from the dataset.*

We see from the histograms in figure 1.1 that the follow-up for death is till end of November 2014, but for renal disease and ESRD which seem to be till sometime in May 2015. The latter is however not usable, because we do not have the deaths occurring between Nov 2014 and May 2015.

The entry and exit dates to SDC seem a bit oddly distributed, and not all persons with an entry date have an exit date, whereas none of those without entry have an exit date:

```
> with( nuf, table( has.in = !is.na(doin),
+                   has.ex = !is.na(dox), exclude=NULL ) )
       has.ex
has.in  FALSE TRUE <NA>
  FALSE   196    0    0
  TRUE   5674 9314    0
  <NA>      0    0    0
> range( nuf$dox, na.rm=TRUE )
[1] 1993.899 2014.901
```

We can explore whether any of the funny patterns in the separatedates are detectable in the joint patterns:

```
> with( nuf, plot( ifelse( doin<1993.754, 1993.5-runif(nrow(nuf)), doin ),
+                  pmin( dox, 2015.3+runif(nrow(nuf)), na.rm=TRUE ),
+                  xlab="Date of entry to SDC",
+                  ylab="Date of exit from SDC",
+                  pch=16, cex=0.3 ) )
> for( i in 0:2 ) abline( i, 1, col="red" )
> rug( 2013+0:2/2, side=2 )
```

From figure 1.2 we see the very prominent exit date of 1 Jan, 1 Jul and 31 Dec 2013. Also we can see the aggregation of entry dates around 2010, as is also apparent from the histogram of entry dates. Finally, we also see that a large fraction of the exit dates are within the firts two years of entry; in the band between the red 45° lines.

1.3.1 Date variable relations

First we provide an overview of the date variables paired, so that we can see to what extent they are in the wrong order. We only plot for 5000 records instead of all 500,000, in order to keep the size of the graph manageable:

```
> dn <- grnam( "^do", nuf )
[1] "dob"    "doDM"   "dodm"   "doin"   "dockd"  "doesrd" "dodth"  "dox"
> par( bty="o" )
> pairs( nuf[,dn], gap=0, pch=16, cex=0.2,
+        panel=function(x,y,...) {points(x,y,...);abline(0,1,col="red")} )
```

1.4 GFR and other renal measurements

We make a brief overview of the number of records per person, as well as the number of GFR, resp EGFR measurements

```
> addmargins( with( nef, table( gfr=!is.na(gfr), egfr=!is.na(egfr) ) ) )
```

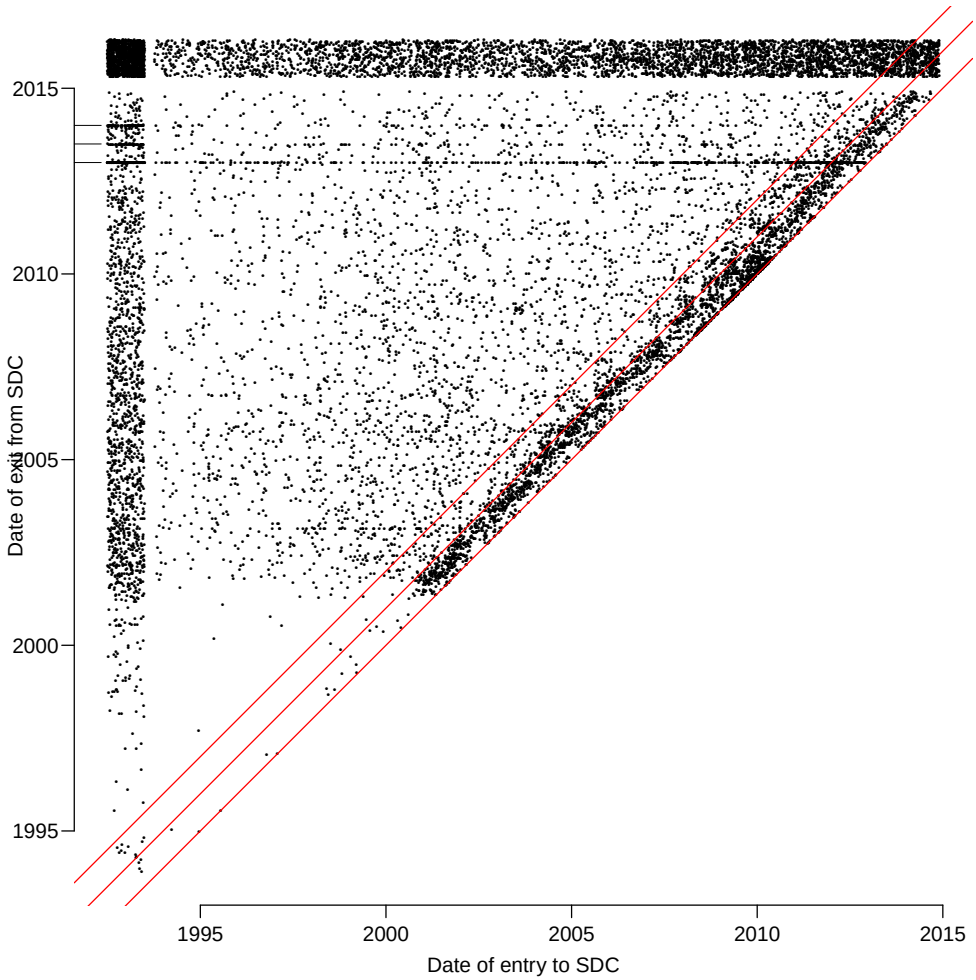

Figure 1.2: *Joint distribution of entry and exit dates to SDC. The band to the left are those with date of entry coded as 1993-10-03, and the band at the top those with date of exit missing.*

```
       egfr
gfr      FALSE   TRUE    Sum
  FALSE 137424 274705 412129
  TRUE     471   4763   5234
  Sum   137895 279468 417363
```

```
> par( mfrow=c(1,2), mar=c(3,3,1,1), mgp=c(3,1,0)/1.6, yaxs="i", las=1, bty="n" )
> nt <- with( nef, table(table(newid)) )
> plot( with( nef, nt ), type="h", lwd=5, xaxt="n", ylim=c(0,500), xlim=c(0,150),
+       ylab="No. persons", xlab="No. records per person" )
> axis(side=1)
> axis(side=1,at=1:25*10,labels=NA)
> nt <- with( subset( nef, !is.na(egfr) | !is.na(gfr) ), table(table(newid)) )
> plot( with( nef, nt ), type="h", lwd=5, xaxt="n", ylim=c(0,500), xlim=c(0,150),
+       ylab="No. persons", xlab="No. records with (e)GFR per person" )
> axis(side=1)
> axis(side=1,at=1:25*10,labels=NA)
> many <- nt[nt>500]
> names( many )
[1] "1" "2" "3" "4" "5" "6" "7"
```

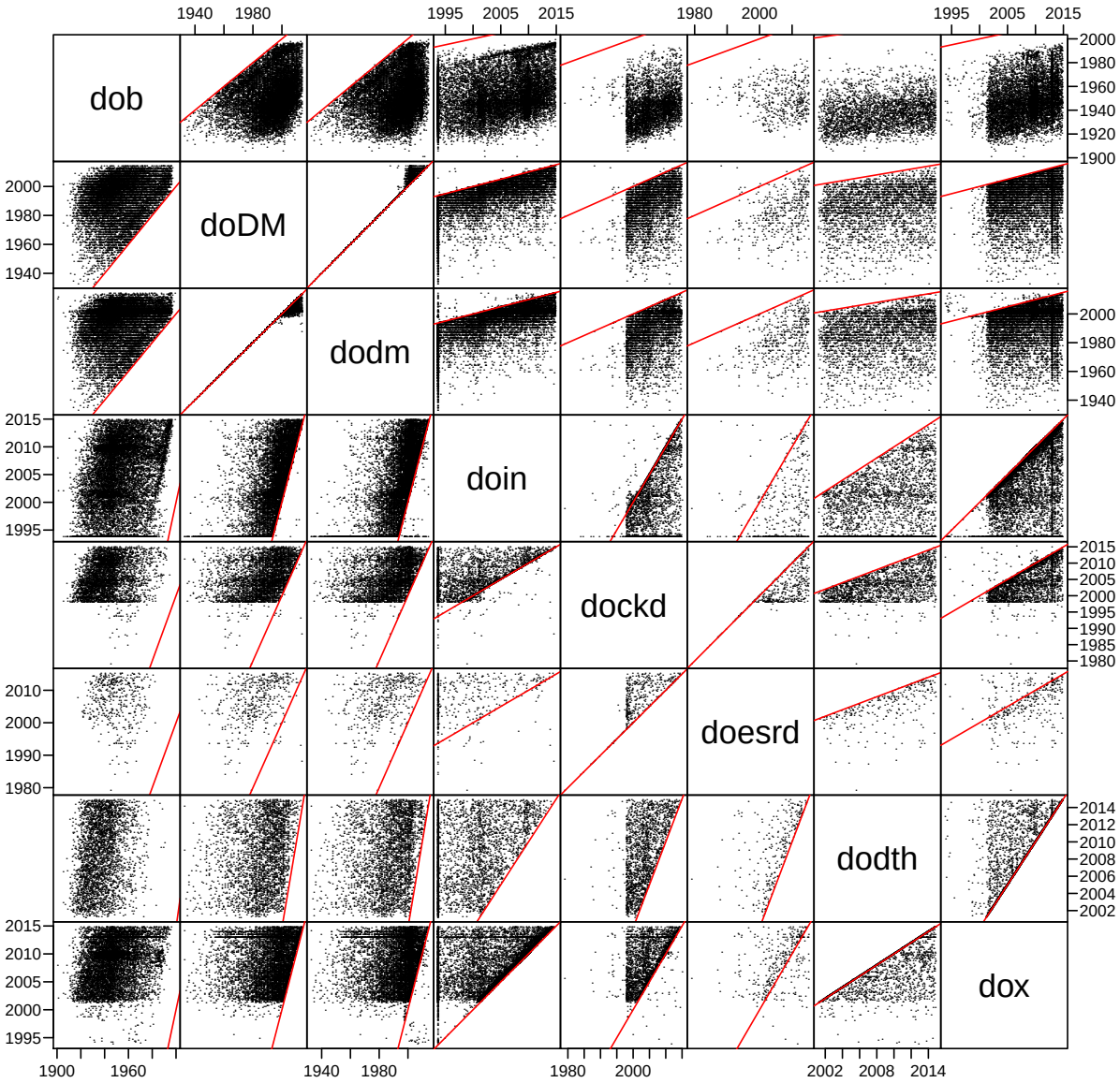

Figure 1.3: *Date variables in the SDC clinical dataset. Each dot represents one person. The red lines are the identity lines, meaning that all points should be on the same side of the lines since the date variables are listed in approximately ascending order.*

```
> for(i in 1:length(many)) text( 10+20*i, 490,
+                                paste(names(many)[i],"\n",many[i]), adj=1 )
```

1.4.1 Renal endpoints

We will be using both `egfr` and `gfr`, as well as `ualbcrea` and `dualb` in the definitions of the renal endpoints:

```
> with( nef, table(eGFR=!is.na(egfr),GFR=!is.na(gfr)) )
```

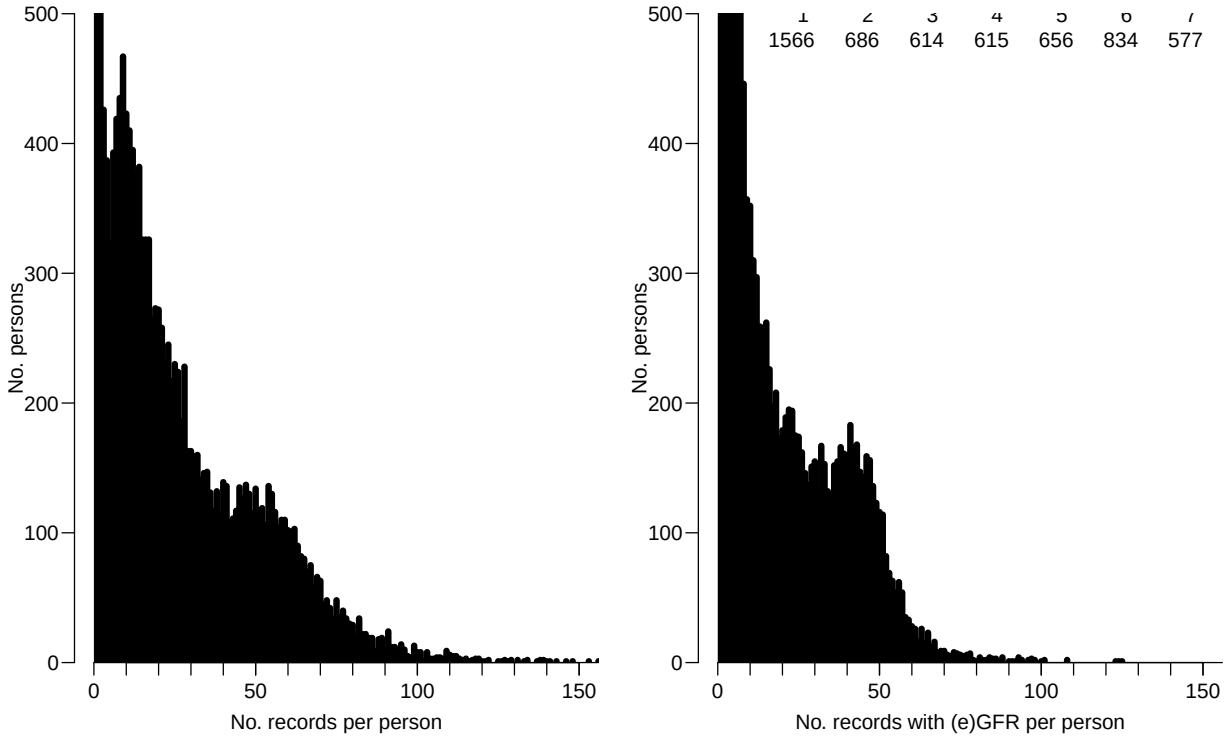

Figure 1.4: *Persons in the database classified by the number of records in the dataset, resp. number of records with GFR or eGFR measurements.*

```
       GFR
eGFR     FALSE   TRUE
  FALSE 137424    471
  TRUE  274705   4763
> with( nef, table(ucr=!is.na(ualbcrea),dualb=!is.na(dualb)) )
       dualb
ucr      FALSE   TRUE
  FALSE 272395  50877
  TRUE   93585    506
```

With this in mind we can define the desired variables from `gfr` and `egfr` and the albumin variabes `dualb` and `ualbcrea`:

```
> nef <- transform( nef, GFR =      pmin( egfr, gfr, na.rm=TRUE ),
+                       ren.st = cut( pmin( egfr, gfr, na.rm=TRUE ),
+                                     breaks=c(0,15,30,45,60,90,Inf),
+                                     include.lowest=TRUE ),
+                       alb.st = cut( pmax(dualb,ualbcrea,na.rm=TRUE),
+                                     breaks=c(0,30,300,Inf),
+                                     right=FALSE ) )
> nef$ckd.st <- Relevel( interaction( nef$ren.st, nef$alb.st ),
+                          list( "CKD 5" = 1+0:2*6,
+                                "CKD 4" = 2+0:2*6,
+                                "CKD 3b"= 3+0:2*6,
+                                "CKD 3a"= 4+0:2*6,
+                                "CKD 2" = 5+1:2*6,
+                                "CKD 1" = 6+1:2*6,
+                                "noCKD" = 5:6 ) )
> non.miss <- function(x) sum(x[-length(x)])
```

```
> with( nef, addmargins( table( alb.st, ren.st, useNA="ifany" ),
+                        FUN=list(list(sum,non.miss),list(sum,non.miss)),
+                      quiet=TRUE ) )[c(1:3,6,4,5),c(1:6,9,7,8)]
           ren.st
alb.st      [0,15] (15,30] (30,45] (45,60] (60,90] (90,Inf] non.miss   <NA>    sum
  [0,30)        22     424    1096    1960   25608    55258    84368   9176  93544
  [30,300)      70     644    1316    1595   12654    14189    30468   6598  37066
  [300,Inf)    167     646     989     882    4107     4634    11425   2933  14358
  non.miss     259    1714    3401    4437   42369    74081   126261  18707 144968
  <NA>         366    1291    2789    3790   41224   104218   153678 118717 272395
  sum          625    3005    6190    8227   83593   178299   279939 137424 417363
> with( nef, print( ftable( ckd.st, alb.st, ren.st, row.vars=1:2 ), z="." ) )
                ren.st [0,15] (15,30] (30,45] (45,60] (60,90] (90,Inf]
ckd.st alb.st
CKD 5  [0,30)              22       .       .       .       .        .
       [30,300)            70       .       .       .       .        .
       [300,Inf)          167       .       .       .       .        .
CKD 4  [0,30)               .     424       .       .       .        .
       [30,300)             .     644       .       .       .        .
       [300,Inf)            .     646       .       .       .        .
CKD 3b [0,30)               .       .    1096       .       .        .
       [30,300)             .       .    1316       .       .        .
       [300,Inf)            .       .     989       .       .        .
CKD 3a [0,30)               .       .       .    1960       .        .
       [30,300)             .       .       .    1595       .        .
       [300,Inf)            .       .       .     882       .        .
CKD 2  [0,30)               .       .       .       .       .        .
       [30,300)             .       .       .       .   12654        .
       [300,Inf)            .       .       .       .    4107        .
CKD 1  [0,30)               .       .       .       .       .        .
       [30,300)             .       .       .       .       .    14189
       [300,Inf)            .       .       .       .       .     4634
noCKD  [0,30)               .       .       .       .   25608    55258
       [30,300)             .       .       .       .       .        .
       [300,Inf)            .       .       .       .       .        .
> with( nef, print( table( ESRD=doesrd<=dolab, ckd.st ), z="." ) )
       ckd.st
ESRD    CKD 5 CKD 4 CKD 3b CKD 3a CKD 2 CKD 1 noCKD
  FALSE    12   494    398    176   570   844   353
  TRUE    238   115     53     26   333   385   246
> any.esrd <- subset( nef, !is.na(doesrd) )
> with( any.esrd, length( unique( newid ) ) )
[1] 478
```

We can then save the dataset in the final analysis form:

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

Chapter 2

Analysis

```
> library( Epi )
> load( file="./data/sdc.Rda" )
> sdc[1:10,c(1:8,56,57,59,60)]
```

	newid	sex	dob	dodm	dodd	doin	dox	dolab	GFR	ren.st	doESRD	ESRD
1	1	M	1953.572	1993	2014.579	2002.691	2003.231	2002.691	NA	<NA>	NA	FALSE
2	1	M	1953.572	1993	2014.579	2002.691	2003.231	2002.787	NA	<NA>	NA	FALSE
3	1	M	1953.572	1993	2014.579	2002.691	2003.231	2002.831	NA	<NA>	NA	FALSE
4	1	M	1953.572	1993	2014.579	2002.691	2003.231	2002.886	NA	<NA>	NA	FALSE
5	1	M	1953.572	1993	2014.579	2002.691	2003.231	2003.020	NA	<NA>	NA	FALSE
6	1	M	1953.572	1993	2014.579	2002.691	2003.231	2003.080	NA	<NA>	NA	FALSE
7	1	M	1953.572	1993	2014.579	2002.691	2003.231	2003.209	NA	<NA>	NA	FALSE
9	3	M	1919.461	1976	2008.827	1993.755	2008.827	1998.010	NA	<NA>	NA	FALSE
10	3	M	1919.461	1976	2008.827	1993.755	2008.827	1998.012	NA	<NA>	NA	FALSE
11	3	M	1919.461	1976	2008.827	1993.755	2008.827	1998.344	NA	<NA>	NA	FALSE

```
> sdc[sdc$newid==8,c(1:8,56,57,59,60)]
```

	newid	sex	dob	dodm	dodd	doin	dox	dolab	GFR	ren.st	doESRD	ESRD
189	8	F	1936.701	1985	2002.226	1993.755	2002.226	1998.015	NA	<NA>	1998.779	TRUE
190	8	F	1936.701	1985	2002.226	1993.755	2002.226	1998.070	NA	<NA>	1998.779	TRUE
191	8	F	1936.701	1985	2002.226	1993.755	2002.226	1998.223	NA	<NA>	1998.779	TRUE
192	8	F	1936.701	1985	2002.226	1993.755	2002.226	1998.245	NA	<NA>	1998.779	TRUE
193	8	F	1936.701	1985	2002.226	1993.755	2002.226	1998.439	NA	<NA>	1998.779	TRUE
194	8	F	1936.701	1985	2002.226	1993.755	2002.226	1998.445	NA	<NA>	1998.779	TRUE
195	8	F	1936.701	1985	2002.226	1993.755	2002.226	1998.760	NA	<NA>	1998.779	TRUE
196	8	F	1936.701	1985	2002.226	1993.755	2002.226	1998.779	15	[0,15]	1998.779	TRUE
197	8	F	1936.701	1985	2002.226	1993.755	2002.226	1999.190	NA	<NA>	1998.779	TRUE
198	8	F	1936.701	1985	2002.226	1993.755	2002.226	1999.603	NA	<NA>	1998.779	TRUE
199	8	F	1936.701	1985	2002.226	1993.755	2002.226	1999.814	NA	<NA>	1998.779	TRUE
200	8	F	1936.701	1985	2002.226	1993.755	2002.226	2000.129	NA	<NA>	1998.779	TRUE
201	8	F	1936.701	1985	2002.226	1993.755	2002.226	2000.140	NA	<NA>	1998.779	TRUE
202	8	F	1936.701	1985	2002.226	1993.755	2002.226	2000.397	NA	<NA>	1998.779	TRUE
203	8	F	1936.701	1985	2002.226	1993.755	2002.226	2000.411	NA	<NA>	1998.779	TRUE
204	8	F	1936.701	1985	2002.226	1993.755	2002.226	2000.794	NA	<NA>	1998.779	TRUE
205	8	F	1936.701	1985	2002.226	1993.755	2002.226	2000.890	NA	<NA>	1998.779	TRUE
206	8	F	1936.701	1985	2002.226	1993.755	2002.226	2001.109	NA	<NA>	1998.779	TRUE
207	8	F	1936.701	1985	2002.226	1993.755	2002.226	2001.314	NA	<NA>	1998.779	TRUE
208	8	F	1936.701	1985	2002.226	1993.755	2002.226	2001.637	NA	<NA>	1998.779	TRUE
209	8	F	1936.701	1985	2002.226	1993.755	2002.226	2001.886	NA	<NA>	1998.779	TRUE

2.1 Outcome data

We need to fish out all records with GFR measurements, and subsequently persons with at least two measurements

```
> sdcR <- subset( sdc, !is.na(GFR) & dolab <= doESRD )
> dim( sdcR )
[1] 1244   62
> addmargins( with( sdcR, table(table(newid)) ) )
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  24
 30  27  14  10  18  10   6   8   9   5   8   6   5   1   3   1   3   1   2   1   1   1   1
> # Fishing out persons with at least 2 measurements
> tt <- table(sdcR$newid)
> over1 <- names( tt[tt>1] )
> sdcR <- subset( sdcR, newid %in% over1 )
> addmargins( with( sdcR, table(table(newid)) ) )
  2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  24  25
 27  14  10  18  10   6   8   9   5   8   6   5   1   3   1   3   1   2   1   1   1   1   1
```

Since we are going to analye GFR as a function of time before ESRD, we will need the time to ESRD, `ttESD` as a separate variable:

```
> sdcR <- transform( sdcR, ttESRD = dolab-doESRD )
> hist( sdcR$ttESRD, col="black", breaks=seq(-20,0,0.5) )
```

2.2 Trajectory analyses with latent classes

The following illustrates the use of the `lcmm` package to fit random effects spline models to the trajectories of those that end with ESRD. Thus we are conditioning on the end state renal disease outcome (`ESRD`), and model how the trajectories of GFR is in these individuals. The purpose of this is to try to identify different *shapes* of GFR-decline up to ESRD.

So we first subset the data to those persons who actually get ESRD. Since `lcmm` does not accept the usual model formulae we must explicitly construct the columns of the spline basis (note that the `Ns` is a wrapper from the `Epi` package to simplify definition of natural splines). Also note that `detrend` is a function from `Epi` that makes a projection of the columns of the spline basis on the orthogonal complement to the constant plus the time variable. The resulting columns are thus the non-linear effects of the time variable, in the case `ttESRD`:

```
> library( lcmm )
> library( splines )
> esrd <- subset( sdcR, ESRD )
> esrd$age <- esrd$dolab - esrd$dob
> with( esrd, round( quantile( ttESRD, 0:10/10 ), 1 ) )
   0%   10%   20%   30%   40%   50%   60%   70%   80%   90%  100%
-16.1 -11.8  -9.7  -7.4  -5.5  -3.8  -2.6  -1.7  -1.0   0.0   0.0
> ( kn <- seq(-15,0,,5) )
[1] -15.00 -11.25  -7.50  -3.75   0.00
> MM <- Ns( esrd$ttESRD, knots=kn )
> MM <- detrend( MM, esrd$ttESRD )
> ( colnames(MM) <- paste("x",colnames(MM),sep="") )
```

```
[1] "x1" "x2" "x3"
> esrd <- cbind( esrd, MM )
> head( MM )
               x1         x2        x3
[1,] -0.268399677 -0.1677629 0.1662526
[2,] -0.129230404 -0.1982983 0.1940913
[3,] -0.030780993 -0.1996719 0.1684513
[4,] -0.030780993 -0.1996719 0.1684513
[5,] -0.030780993 -0.1996719 0.1684513
[6,]  0.002985959 -0.1977835 0.1554466
```

We have now set up data to fit the model; the columns `x1`, `x2`, `x3` and `x4` represent the non-linear effects of time before ESRD. This means that that coefficient to `ttESRD` represents the *average* time trend in eGFR over time. Thus it is possible to compare the size of this with the sd of the random effects (that is the between person variation in slopes). The argument `nwg=TRUE` scales the random-effect covariance between classes:

```
> system.time(
+ fitspl <- hlme( GFR ~ x1 + x2 + x3 + ttESRD + age + sex,
+           mixture = ~ x1 + x2 + x3 + ttESRD,
+            random = ~ ttESRD,
+           subject = 'newid', ng = 3, nwg=TRUE, data = esrd ) )
Be patient, hlme is running ...
The program took 52.08 seconds
   user  system elapsed
 52.090   0.001  52.084
> fitspl
Heterogenous linear mixed model
     fitted by maximum likelihood method

hlme(fixed = GFR ~ x1 + x2 + x3 + ttESRD + age + sex, mixture = ~x1 +
    x2 + x3 + ttESRD, random = ~ttESRD, subject = "newid", ng = 3,
    nwg = TRUE, data = esrd)

Statistical Model:
     Dataset: esrd
     Number of subjects: 148
     Number of observations: 1214
     Number of latent classes: 3
     Number of parameters: 25

Iteration process:
     Convergence criteria satisfied
     Number of iterations:  24
     Convergence criteria: parameters= 5.1e-08
                         : likelihood= 8.8e-11
                         : second derivatives= 1.7e-15

Goodness-of-fit statistics:
     maximum log-likelihood: -4238.2
     AIC: 8526.41
     BIC: 8601.34
> summary( fitspl )
Heterogenous linear mixed model
     fitted by maximum likelihood method
```

```
hlme(fixed = GFR ~ x1 + x2 + x3 + ttESRD + age + sex, mixture = ~x1 +
    x2 + x3 + ttESRD, random = ~ttESRD, subject = "newid", ng = 3,
    nwg = TRUE, data = esrd)

Statistical Model:
     Dataset: esrd
     Number of subjects: 148
     Number of observations: 1214
     Number of latent classes: 3
     Number of parameters: 25

Iteration process:
     Convergence criteria satisfied
     Number of iterations:  24
     Convergence criteria: parameters= 5.1e-08
                         : likelihood= 8.8e-11
                         : second derivatives= 1.7e-15

Goodness-of-fit statistics:
     maximum log-likelihood: -4238.2
     AIC: 8526.41
     BIC: 8601.34


Maximum Likelihood Estimates:

Fixed effects in the class-membership model:
(the class of reference is the last class)

                     coef      Se   Wald p-value
intercept class1  1.48259 0.27612  5.369 0.00000
intercept class2 -0.30545 0.47927 -0.637 0.52392

Fixed effects in the longitudinal model:

                      coef       Se    Wald p-value
intercept class1  14.47848  1.99227   7.267 0.00000
intercept class2  15.54789  1.95576   7.950 0.00000
intercept class3  26.53283  2.67008   9.937 0.00000
x1 class1         -4.02721  1.81554  -2.218 0.02654
x1 class2        -39.28702  6.05635  -6.487 0.00000
x1 class3          3.77527  6.05364   0.624 0.53287
x2 class1         -2.05853  1.53305  -1.343 0.17935
x2 class2         16.66596  4.22654   3.943 0.00008
x2 class3         30.43634  6.54339   4.651 0.00000
x3 class1         -7.71538  4.13201  -1.867 0.06187
x3 class2         43.85254 15.06561   2.911 0.00361
x3 class3          3.91147 17.12475   0.228 0.81933
ttESRD class1     -4.61879  0.27289 -16.925 0.00000
ttESRD class2     -6.18769  0.93584  -6.612 0.00000
ttESRD class3    -10.03379  1.74118  -5.763 0.00000
age               -0.01017  0.02996  -0.340 0.73419
sexF              -1.41796  0.74607  -1.901 0.05736


Variance-covariance matrix of the random-effects:
          intercept   ttESRD
intercept  17.84505
ttESRD     10.28227 43.57987
```

```
                                    coef         se
Proportional coefficient class1 0.3106424 0.06732407
Proportional coefficient class2 0.3318239 0.15959066
Residual standard error:        6.2460811 0.14336491
```

Once we fitted the model we can have a look at how the posterior probabilities of being in the assigned classes look:

```
> postprob( fitspl )
Posterior classification:
  class1 class2 class3
N    111  15.00  22.00
%     75  10.14  14.86

Posterior classification table:
     --> mean of posterior probabilities in each class
        prob1  prob2  prob3
class1 0.9329 0.0451 0.0219
class2 0.1260 0.7790 0.0950
class3 0.0317 0.0483 0.9200

Posterior probalities above a threshold (%):
         class1 class2 class3
prob>0.7  95.50  60.00  90.91
prob>0.8  88.29  53.33  81.82
prob>0.9  78.38  40.00  81.82
> names( fitspl )
 [1] "ns"        "ng"        "idea0"     "idprob0"   "idg0"      "idcor0"    "loglik"    "b
[10] "gconv"     "conv"      "call"      "niter"     "dataset"   "N"         "idiag"     "p
[19] "predRE"    "Xnames"    "Xnames2"   "cholesky"  "na.action"
> str( ppr <- fitspl$pprob )
'data.frame':        148 obs. of  5 variables:
 $ newid: num  60 222 283 406 411 440 507 530 670 709 ...
 $ class: int  1 1 1 1 2 3 2 1 2 3 ...
 $ prob1: num  1 1 0.918 0.598 0.243 ...
 $ prob2: num  8.42e-14 3.19e-12 3.40e-02 3.63e-01 6.09e-01 ...
 $ prob3: num  6.68e-08 1.45e-04 4.85e-02 3.86e-02 1.48e-01 ...
> ncl <- table( ppr$class )
> clr <- c("red","black","blue","limegreen")
> par( mfrow=c(1,3), mar=c(3,0,1,1), oma=c(0,3,0,0), las=1, bty="n", mgp=c(3,1,0)/1.6 )
> for( i in 1:3 )
+    {
+    hist( ppr[ppr$class==i,i+2], breaks=0:50/50,
+          col=clr[i], border=clr[i], ylim=c(0,60), main="",
+          xlab="", yaxt="n", yaxs="i" )
+    if( i==1 ) axis( side=2 )
+    text( 0.1, 50, paste( "Class", i, ": n=", ncl[i] ), font=2,
+         col=clr[i], cex=1.5, adj=0 )
+    }
```

A slightly more informative plot of the posterior probabilities is obtained by looking at the pairwise

```
> par( bty="o")
> pairs( ppr[,2+1:3], pch=16, col=clr[ppr$class], cex=1.5, gap=0 )
```

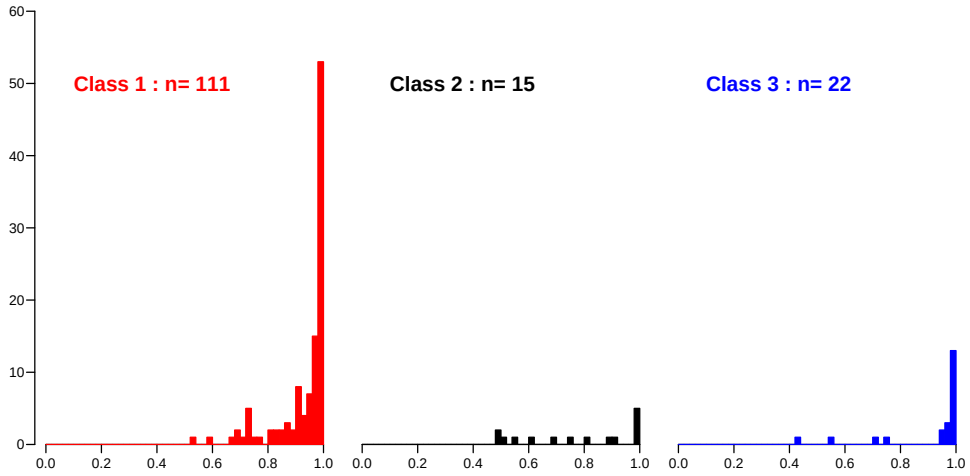

Figure 2.1: *Posterior probabilities of class membership for the ESRD cases modelled.*

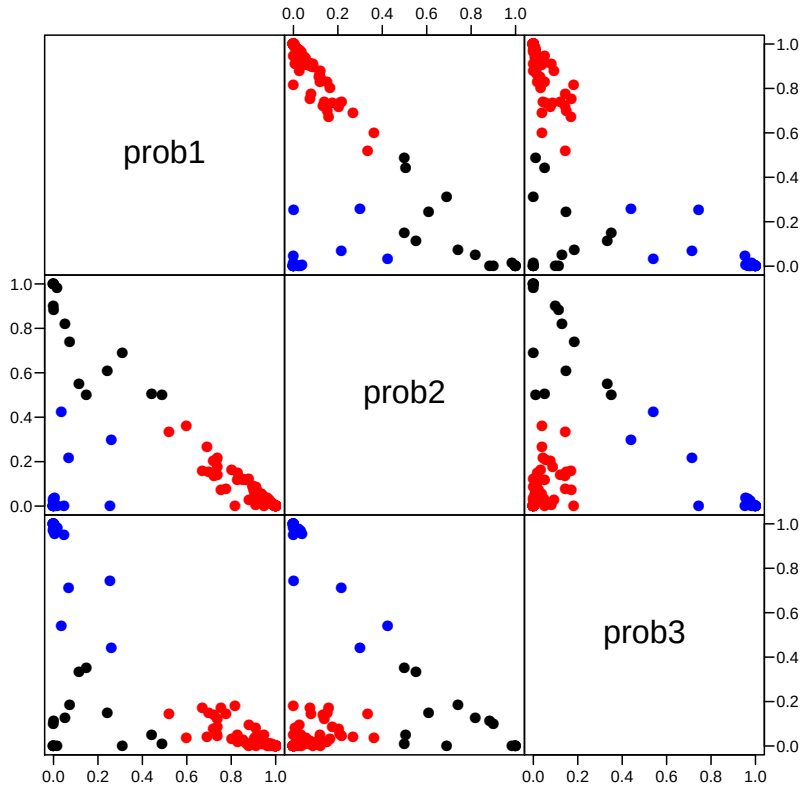

Figure 2.2: *Pairwise posterior probabilities. It is seen that class 1 (red) is not well discriminated from class 2 (black; the largest class)*

In order to plot the estimated trajectories we extract a prediction data frame from the analysis data frame. This is necessary because the construction of the de-trended version of the variables depends on data. Incidentally, also the

```
> wh <- match( sort(unique(esrd$ttESRD)), esrd$ttESRD )
> plotdata <- data.frame(1,MM[wh,],esrd$ttESRD[wh],
```

```
+                                60+esrd$ttESRD[wh],
+                           sex=factor("M",levels=c("F","M")) )
> names(plotdata)[-7] <- fitspl$Xnames[-7]
> par( mar=c(3,3,1,1), mgp=c(3,1,0)/1.6 )
> outspl <- plot.predict.hlme(fitspl,plotdata,var.time="ttESRD",
+                             legend.loc="topright",col=clr,lwd=4,
+                             ylim=c(0,160), main="")
> str( outspl )
'data.frame':        865 obs. of  10 variables:
 $ time        : num  -16.1 -15.9 -15.7 -15.6 -15.1 ...
 $ class1      : num  92.7 91.5 90.1 89.5 86.3 ...
 $ class2      : num  110 110 110 110 110 ...
 $ class3      : num  187 185 182 181 175 ...
 $ lower.class1: num  83.1 82.1 80.9 80.5 77.9 ...
 $ lower.class2: num  80.6 81.3 82.1 82.3 83.9 ...
 $ lower.class3: num  131 130 128 127 123 ...
 $ upper.class1: num  102.4 100.9 99.2 98.5 94.8 ...
 $ upper.class2: num  139 139 138 138 137 ...
 $ upper.class3: num  244 240 237 235 227 ...
> # datasub <- merge(datasub, fitspl$pprob[,1:2], by ="id")
```

2.2.1 2 do next

The latent class trajectory distribution of the persons with event is just a rough guide to the possible patterns; and it would be useful to compare this to the general pattern in GFR among those without event (so far). Hence we will:

- fit separate random effects models to the subgroups identified, allowing us to get estimates of the between-person variation, in order to assess to which extent the variation between persons in different classes is the same.

  We shall use the folowing type of random effects model for GFR-measurement y_{it}:

  $$y_{it} = f(t) + \alpha_a + \alpha_s + a_i + b_{it} + e_{it}$$

- extend the models with random slopes and see how these vary between persons.
- plot individual observed trajectories for different classes
- fit a model with linear effect of the time for the patients without ESRD.
- fit a common model for all patients using current age and duration of diabetes as predictors of GFR in addition to sex.
- extend this model with clinical *baseline* variables
- extend this model with *updated* clinical variables

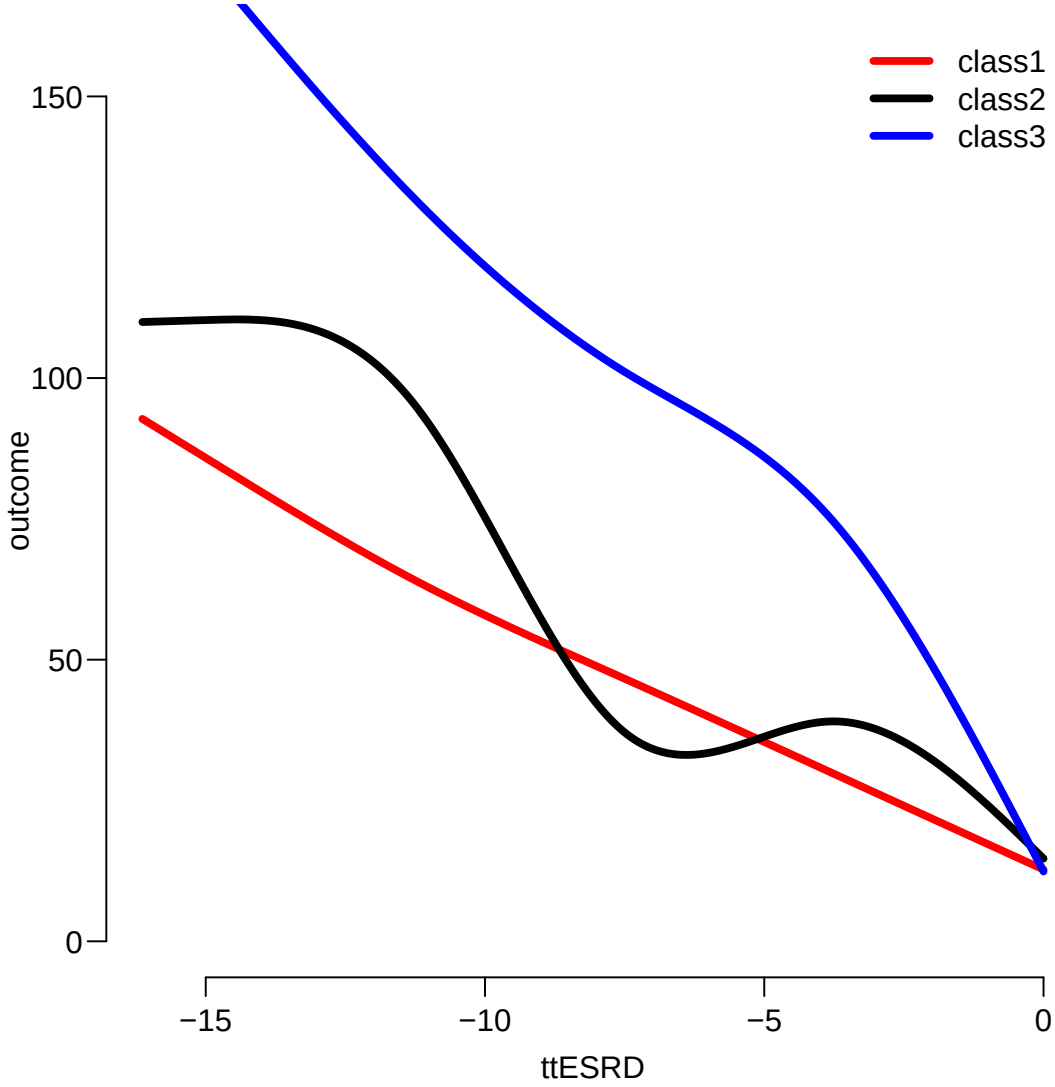

Figure 2.3: *Predicted mean trajectories for the three latent classes of persons developing ESRD. Note that the number of persons in the classes as derived are quite unevenly distributed, the classes have 19, 122 and 10 persons in the classes.*