

# Clinical nephropathy from SDC

---

SDC

August 2015

<http://bendixcarstensen.com/>

Version 3

Compiled Wednesday 14<sup>th</sup> September, 2016, 16:15  
from: /home/bendix/sdc/proj/HKPWH/SDC.tex

Bendix Carstensen   Steno Diabetes Center, Gentofte, Denmark  
& Department of Biostatistics, University of Copenhagen  
[bxo@steno.dk](mailto:bxo@steno.dk)  
<http://BendixCarstensen.com>

Dorte Vistisen   Steno Diabetes Center, Gentofte, Denmark  
[dtvs@steno.dk](mailto:dtvs@steno.dk)

# Contents

<b>1</b>	<b>Reading data</b>	<b>1</b>
1.1	Reading the SDC clinical data . . . . .	1
1.1.1	The <b>Stata</b> dataset . . . . .	1
1.2	Dates and events . . . . .	3
1.2.1	Date problems . . . . .	4
1.3	Overview of dates . . . . .	13
1.3.1	Date variable relations . . . . .	15
1.4	GFR and other renal measurements . . . . .	15
1.4.1	Renal endpoints . . . . .	17
<b>2</b>	<b>Descriptives</b>	<b>20</b>
2.1	Date variables . . . . .	20
2.2	Data overview . . . . .	20
<b>3</b>	<b>Analyses</b>	<b>27</b>
3.1	Outcome data . . . . .	27
3.2	Trajectory analyses with latent classes . . . . .	28

# Chapter 1

## Reading 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 dataset 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/nephrohkwkworkdata.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 )

abdominalomfang      units
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"
middelbladglukoseepj "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
147550	2011.925	NA			NA <14	Genstande/uge	NA	0		NA	NA		
147550					NA	NA	NA	NA		0	NA	NA	NA
147550					NA	NA	NA	NA		NA	0		
147550					NA	Denmark		NA	NA		NA	NA	
147550					NA	NA							

```
> 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 original 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 measurements 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 constitutes 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:

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

and for the disease outcomes:

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

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 diastoliskepjd diurese dualb duplicates egfr gad gfr
16886          NA          NA          NA          NA          NA    0    NA    NA    NA
16887          NA          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          NA    0
16887          NA    NA    NA    NA          NA    NA          NA          NA    0
      pcreatinin region      rygning systoliskepjd trans triglycerid tsh ualbcrea vldl
16886          NA Denmark          Ikke ryger          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 diastoliskepjd diurese dualb duplicates egfr gad gfr
72865          NA          NA          NA          NA          NA    0    NA    NA    NA
72866          NA          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 systoliskepjd 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 diastoliskepjd diurese dualb duplicates egfr gad gfr
104323          NA          NA          NA          NA          NA    0    NA    NA    NA
104324          NA          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          NA    0
104324          NA    NA    NA    NA          NA    NA          NA          NA    0
      pcreatinin region      rygning systoliskepjd 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	dmttype	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	dmttype	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	dmttype	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	diastoliskep	diurese	dualb	duplicates				
233690	NA	NA	D\xf8d	NA	NA	NA	NA	NA	0			
233691	NA	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	systoliskep	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	dmttype	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	NA		
243772	2005.926	2001.921	2002.69		NA	NA	NA	NA	NA	NA		
243777	2005.926	2001.921	2002.69		NA	NA	0	NA	NA	NA		
	civilstandskode	cpeptid	diastoliskep	diurese	dualb	duplicates	egfr	gad	gfr			
243771		NA	NA	NA	NA	NA	0	NA	NA	NA		
243772		NA	NA	NA	NA	NA	0	NA	NA	NA		
243777		NA	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	NA			
243772	NA	NA	NA	NA	NA	NA	NA	NA	NA			
243777	NA	NA	NA	NA	NA	NA	NA	0	NA			
	pcreatinin	region	rygning	systoliskep	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	dmttype	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	NA		
263669	2001.439	1993.754	1998.742		NA	NA	NA	NA	NA	NA		
263671	2001.439	1993.754	1998.742		NA	NA	NA	NA	NA	NA		
	civilstandskode	cpeptid	diastoliskep	diurese	dualb	duplicates	egfr	gad	gfr			
263668		NA	NA	NA	NA	NA	0	NA	NA	NA		
263669		NA	NA	NA	NA	NA	0	NA	NA	NA		
263671		NA	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	NA			
263669	NA	NA	NA	NA	NA	NA	NA	NA	NA			
263671	NA	NA	NA	NA	NA	NA	NA	NA	NA			
	pcreatinin	region	rygning	systoliskep	trans	triglycerid	tsh	ualbcrea	vldl	weight		
263668	NA	<NA>		NA	NA	NA	NA	NA	NA	NA		



```

      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=2000:2016)
> hh( nuf$dockd , "Date of CKD" , breaks=seq(1979,2016,1/ 2) ) ; axis(side=1,at=1979:2016)
> hh( nuf$doesrd , "Date of ESRD" , breaks=seq(1979,2016,1/ 2) ) ; axis(side=1,at=1979:2016)
> hh( nuf$dodth , "Date of death" , breaks=seq(2000,2016,1/12) ) ; abline(v=2000:2016)
> hh( nuf$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

```

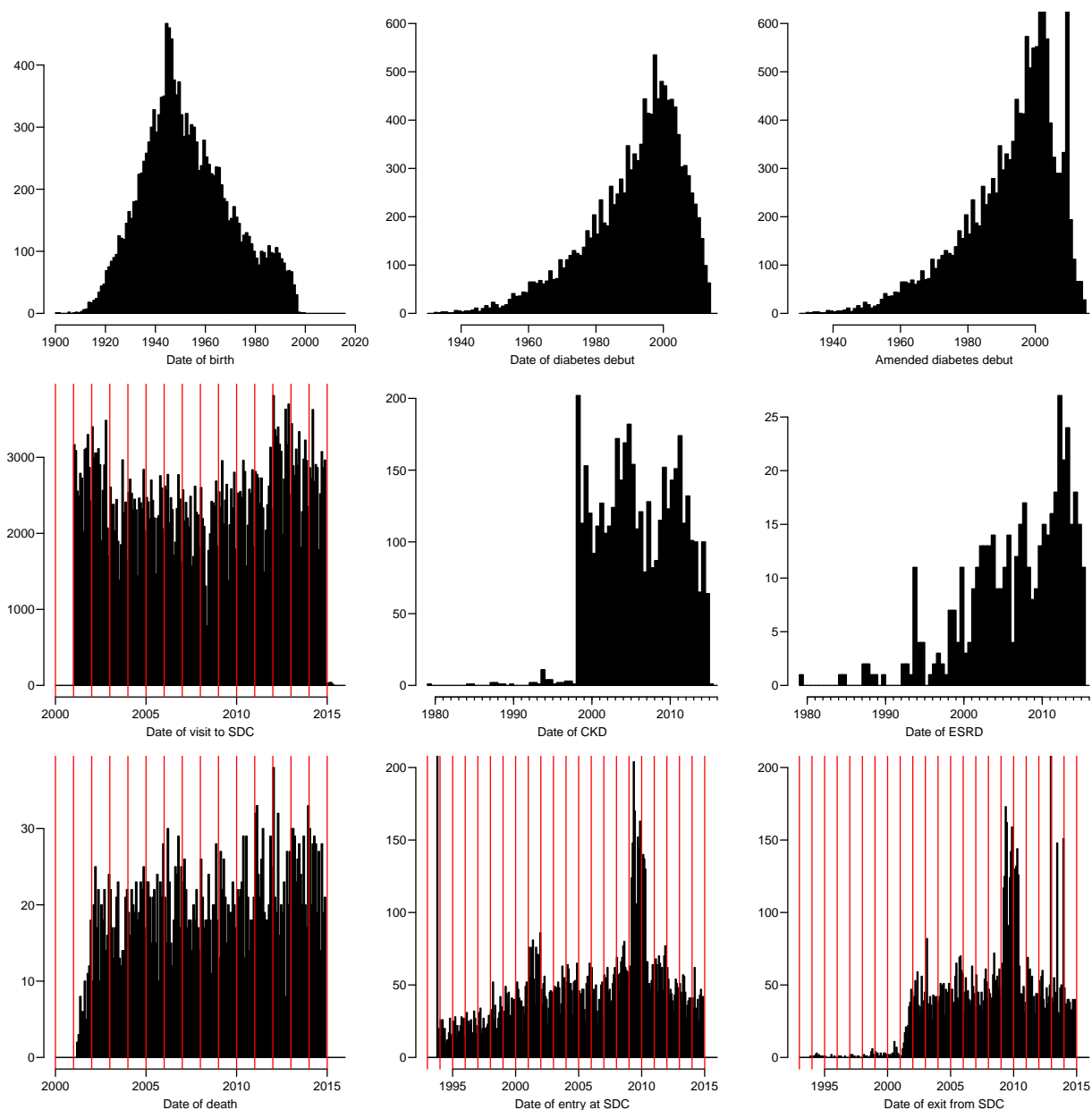


Figure 1.1: Histograms of various dates from the dataset.

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 first 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) ) ) )
```

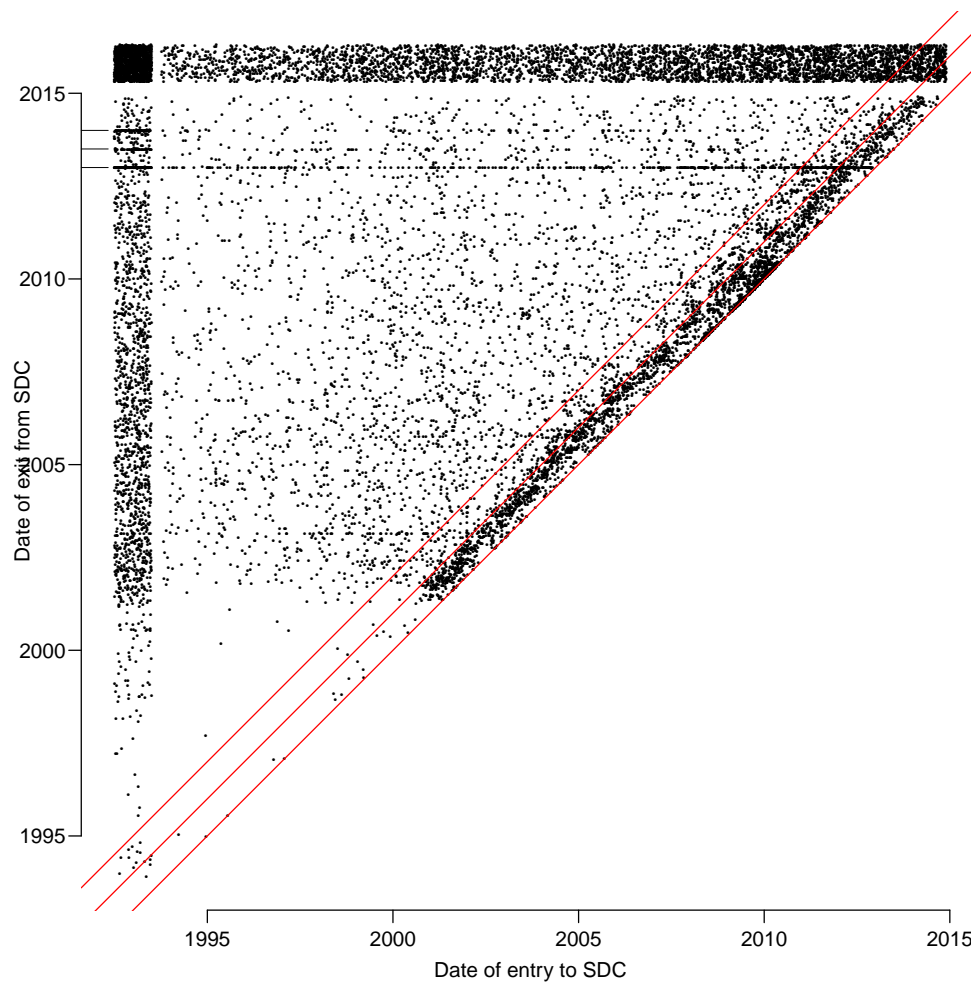


Figure 1.2: Joint distribution of entry and exit dates to SDC. The band to the left are those with date of entry coded as 1993-10-03, and the band at the top those with date of exit missing.

```

      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"

```

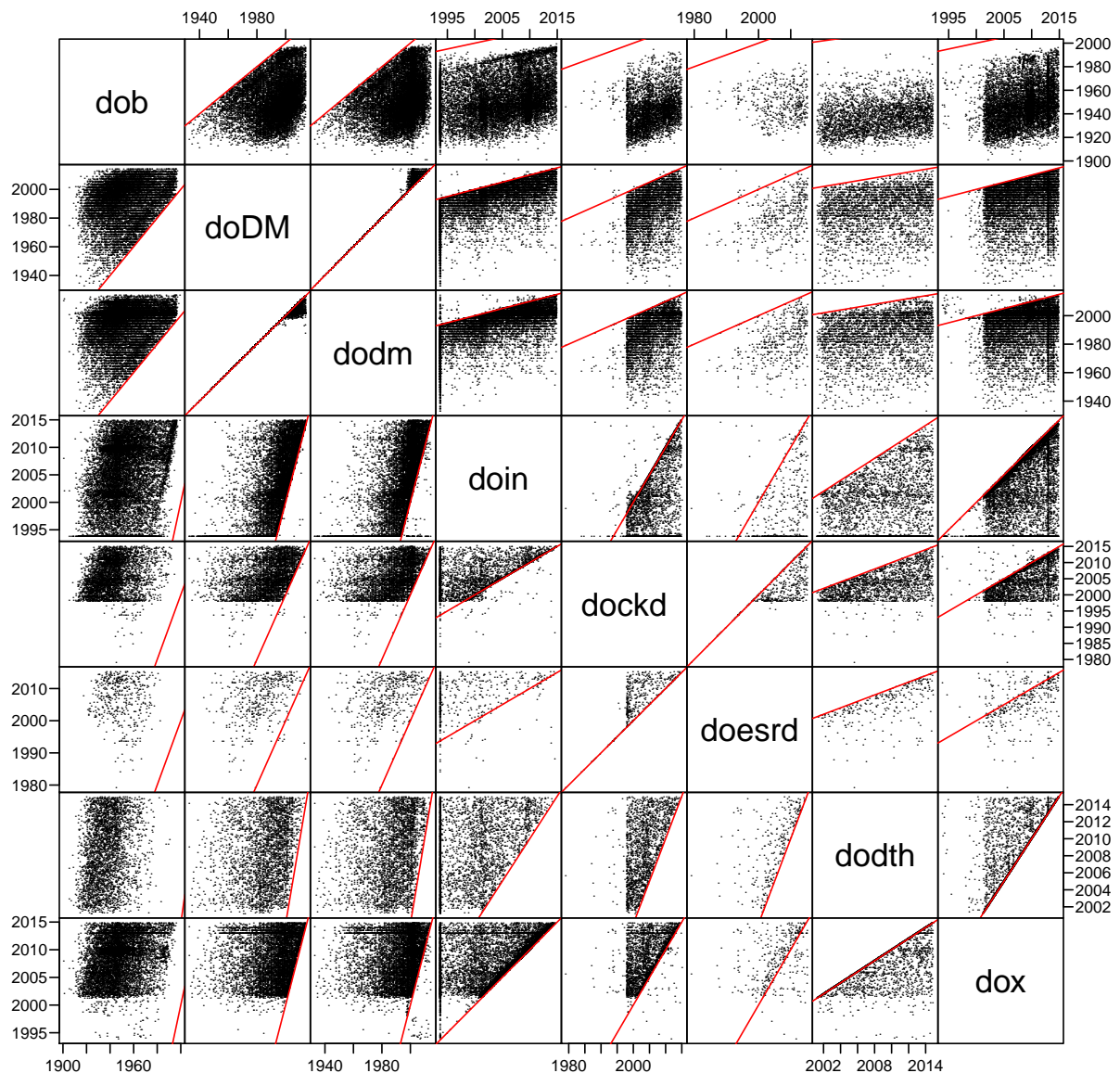


Figure 1.3: Date variables in the SDC clinical dataset. Each dot represents one person. The red lines are the identity lines, meaning that all points should be on the same side of the lines since the date variables are listed in approximately ascending order.

```
> 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)) )
```

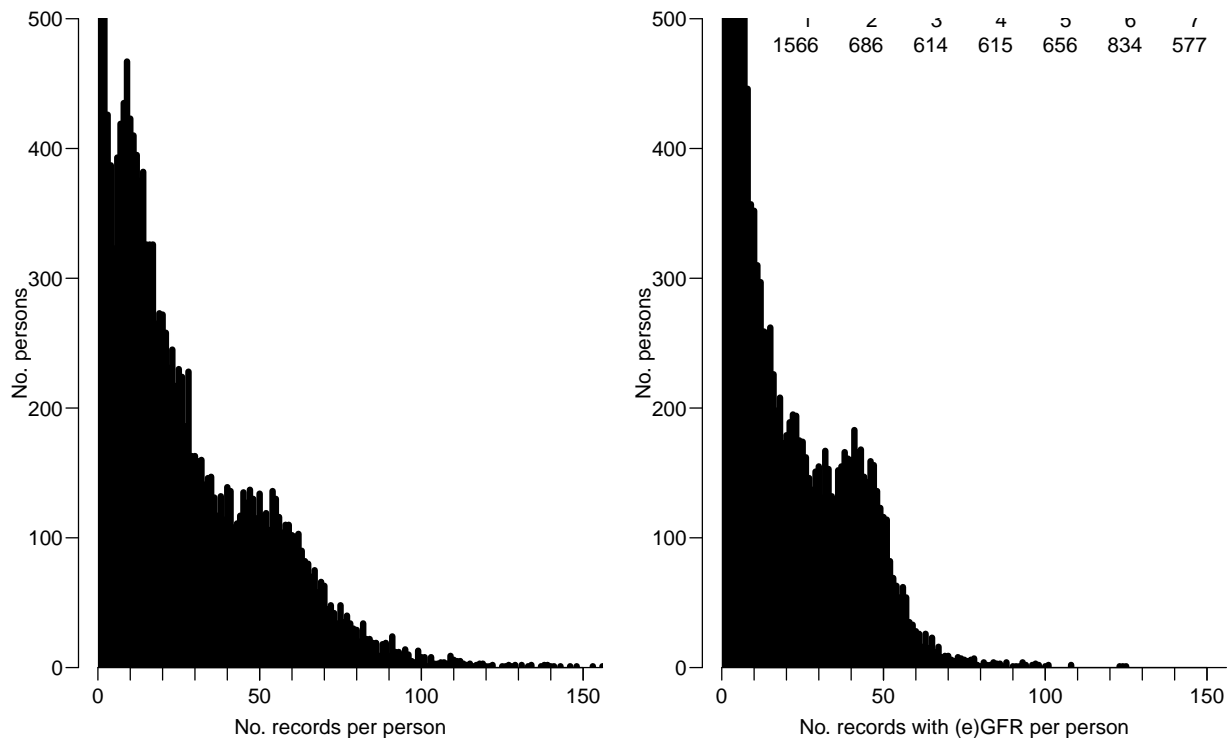


Figure 1.4: Persons in the database classified by the number of records in the dataset, resp. number of records with GFR or eGFR measurements.

```

      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 variables `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="." ) )

```

ckd.st	alb.st	ren.st	[0,15]	(15,30]	(30,45]	(45,60]	(60,90]	(90,Inf]
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

## Descriptives

### 2.1 Date variables

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:

```
> load( file="./data/sdc.Rda" )
> ( dn <- grep("do",names(sdc)) )
[1] 3 4 5 6 7 8 10 53 59 61
> names(sdc)[dn]
[1] "dob"          "dodm"          "dodd"          "doin"
[5] "dox"          "dolab"         "abdominalomfang" "sygdom.diabetes"
[9] "doESRD"       "doCKD"
> par( bty="o" )
> pairs( sdc[sample(1:nrow(sdc),5000),dn[c(1,2,4,6,10,9,5,3)]], gap=0, pch=16, cex=0.2,
+       panel=function(x,y,...) {points(x,y,...);abline(0,1,col="red")})
```

### 2.2 Data overview

First we just show the first few records of the data frame:

```
> head( sdc )
  newid sex    dob dodm    dodd    doin    dox    dolab alat abdominalomfang
1     1  M 1953.572 1993 2014.579 2002.691 2003.231 2002.691    NA              NA
2     1  M 1953.572 1993 2014.579 2002.691 2003.231 2002.787    NA              NA
3     1  M 1953.572 1993 2014.579 2002.691 2003.231 2002.831    NA              NA
4     1  M 1953.572 1993 2014.579 2002.691 2003.231 2002.886    NA              NA
5     1  M 1953.572 1993 2014.579 2002.691 2003.231 2003.020    NA              NA
6     1  M 1953.572 1993 2014.579 2002.691 2003.231 2003.080    NA              NA
  alkohol asat bas cpeptid duna d.vitamin diastoliskepj diastoliskhjemme diurese dualb fe
1      NA   NA   NA     NA   NA      NA             NA              NA   NA   NA NA NA
2      28 177    NA    787   NA      NA             93              NA   NA   NA NA 18
3      NA   NA   NA     NA   NA      NA             95              NA   NA   NA NA NA
4      24 159    NA     NA   NA      NA             NA              NA   NA   NA NA NA
5      NA   NA   NA     NA   NA      NA             85              NA   NA   NA NA NA
6      27 152    NA     NA   NA      NA             92              NA   NA   NA NA NA
  feratio gad gfr gluc hdl hvilepuls kalium ldl natrium pcreatinin systoliskepj
1      NA   NA   NA   NA   NA      NA      NA   NA   NA      NA              NA
```

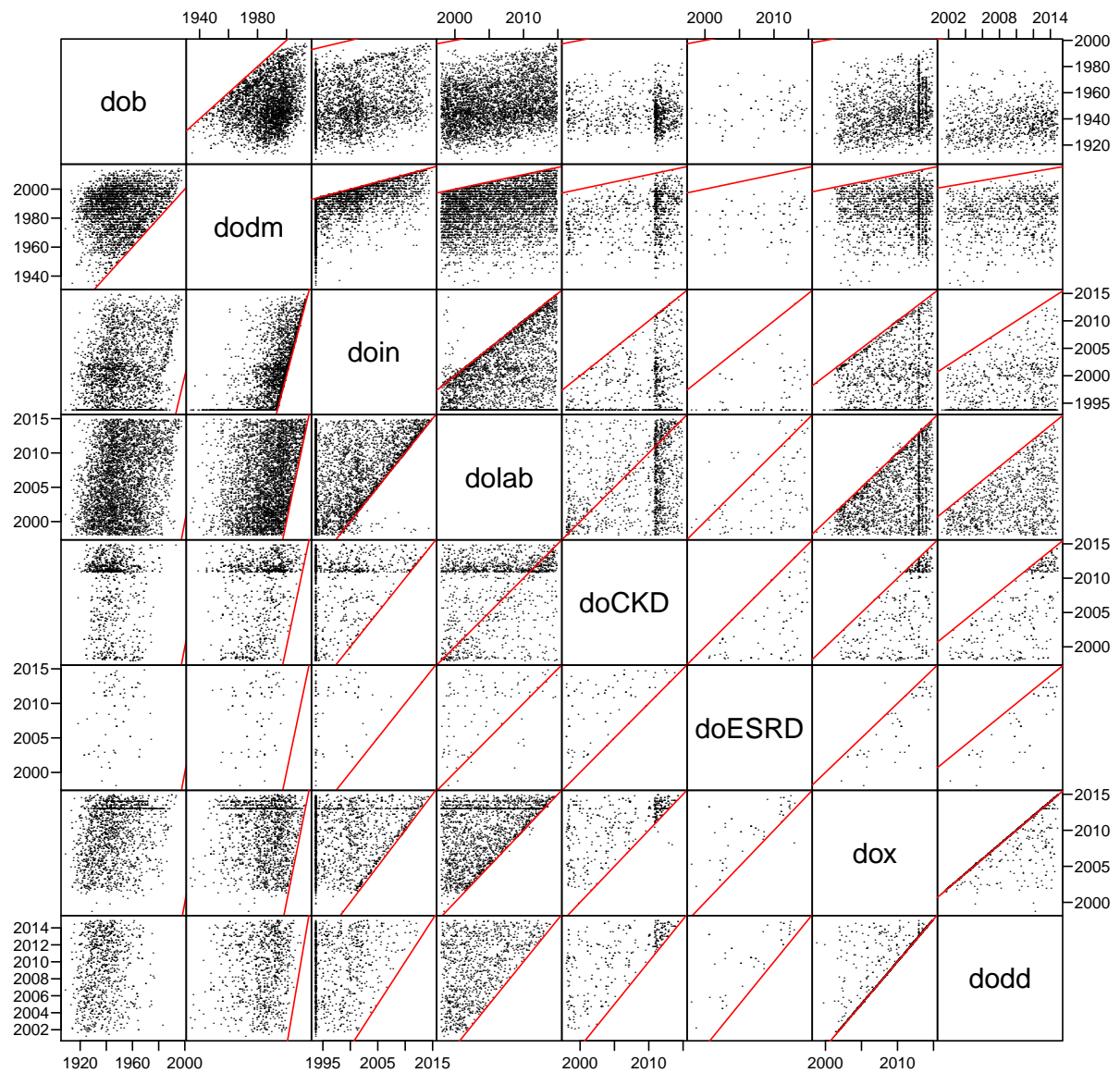


Figure 2.1: Date variables in the SDC clinical dataset. The red lines are the identity lines, meaning that all points should be on the same side of the lines since the date variables are listed in ascending order.

2	22	NA	NA	NA	1.30	NA	3.8	3.2	136	82	143		
3	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	150		
4	NA	NA	NA	NA	1.25	NA	NA	2.6	NA	NA	NA		
5	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	140		
6	NA	NA	NA	NA	1.21	NA	3.9	NA	140	87	144		
	systolik	hjemme	tchol	tsh	trans	triglycerid	ualb	crea	vldl	.merge	b12	blodglukose	bmi
1		NA	NA	NA	NA	NA	4	NA	1	NA	NA	NA	30
2		NA	5.2	2	41	1.45	NA	0.7	1	NA	NA	NA	NA
3		NA	NA	NA	NA	NA	NA	NA	1	NA	NA	NA	NA
4		NA	4.2	NA	NA	0.85	NA	0.4	1	NA	NA	NA	NA
5		NA	NA	NA	NA	NA	NA	NA	1	NA	NA	NA	NA
6		NA	3.7	NA	NA	1.06	4	0.5	1	NA	NA	NA	NA

```

civilstandskode diastolisk.arm egfr haemoglobin hba1c height middelblodglukoseepj
1 D\xf8d NA NA NA 63 1.79 NA
2 D\xf8d NA NA 10.2 57 NA NA
3 D\xf8d NA NA NA NA NA NA
4 D\xf8d NA NA NA 53 NA NA
5 D\xf8d NA NA NA 51 NA NA
6 D\xf8d NA NA NA 48 NA NA
motion rygning sygdom.diabetes systolisk.arm weight GFR ren.st alb.st doESRD ESRD
1 DE11 NA 97.3 NA <NA> [0,30) NA FALSE
2 DE11 NA 99.7 NA <NA> <NA> NA FALSE
3 DE11 NA 96.5 NA <NA> <NA> NA FALSE
4 DE11 NA 94.3 NA <NA> <NA> NA FALSE
5 DE11 NA NA NA <NA> <NA> NA FALSE
6 DE11 NA 96.6 NA <NA> [0,30) NA FALSE
doCKD CKD
1 NA FALSE
2 NA FALSE
3 NA FALSE
4 NA FALSE
5 NA FALSE
6 NA FALSE

```

We then provide an overview over how many visits (well,dates) there are per person:

```

> par( mar=c(3,3,1,1), mgp=c(3,1,0)/1.6, bty="n" )
> tt <- with( sdc, table(table(newid)) )
> plot( as.numeric(names(tt)), tt, type="h", lwd=3, xaxs="i", xlim=c(0,150.5),
+       xlab="No. of dates", ylab="No. of persons", yaxt="n" )
> abline( v=5.5, col="red" )
> axis( side=2 )

```

It might however be more illuminating to see how many visits with recordings of central clinical variables that are available:

```

> data.frame( 1:ncol(sdc), names(sdc) )
  X1.ncol.sdc. names.sdc.
1           1        newid
2           2          sex
3           3          dob
4           4         dodm
5           5         dodd
6           6         doin
7           7          dox
8           8         dolab
9           9          alat
10          10 abdominalomfang
11          11         alkohol
12          12          asat
13          13          bas
14          14         cpeptid
15          15          duna
16          16         d.vitamin
17          17        diastoliskepj
18          18        diastoliskhjemme
19          19         diurese
20          20         dualb
21          21          fe
22          22        feratio
23          23          gad

```

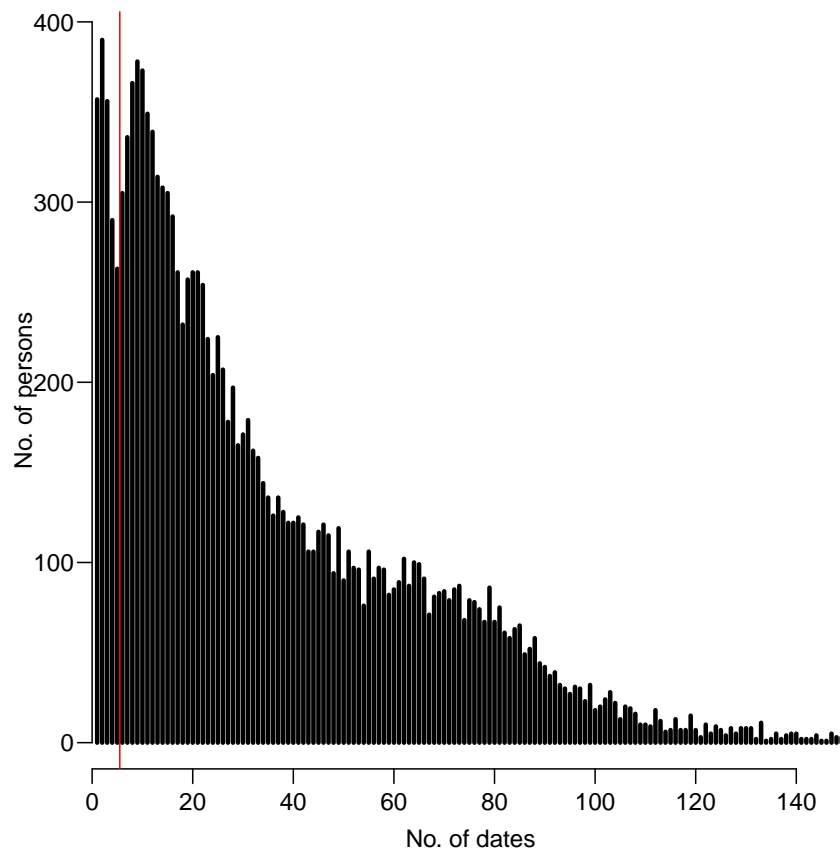


Figure 2.2: Number of recorded dates for patients in the extract — that is patients distributed according to number of records in the dataset. The vertical line is between 5 and 6 visits.

24	24	gfr
25	25	gluc
26	26	hdl
27	27	hvillepuls
28	28	kalium
29	29	ldl
30	30	natrium
31	31	pcreatinin
32	32	systoliskepj
33	33	systoliskhjemme
34	34	tchol
35	35	tsh
36	36	trans
37	37	triglycerid
38	38	ualbcrea
39	39	vldl
40	40	.merge
41	41	b12
42	42	blodglukose
43	43	bmi
44	44	civilstandskode
45	45	diastolisk.arm
46	46	egfr
47	47	haemoglobin
48	48	hba1c

```

49          49          height
50          50 middelblodglukoseepj
51          51          motion
52          52          rygning
53          53          sygdom.diabetes
54          54          systolisk.arm
55          55          weight
56          56          GFR
57          57          ren.st
58          58          alb.st
59          59          doESRD
60          60          ESRD
61          61          doCKD
62          62          CKD

```

```

> wh <- c(
+ "abdominalomfang",
+ "bmi",
+ "blodglukose",
+ "middelblodglukoseepj",
+ "gluc",
+ "hba1c",
+
+ "diastoliskepj",
+ "diastoliskhjemme",
+ "diastolisk.arm",
+ "systoliskepj",
+ "systoliskhjemme",
+ "systolisk.arm",
+
+ "tchol",
+ "hdl",
+ "ldl",
+ "vldl",
+ "triglycerid",
+ "cpeptid",
+
+ "gfr",
+ "egfr",
+ "dualb",
+ "ualbcrea")
> nval <- NArray( list( var=wh, c("no. pers","median no. values") ) )

> n.vis <-
+ function(vn)
+ {
+   df <- sdc[,c("newid",vn)]
+   df <- df[complete.cases(df),]
+   tt <- table(table(df$newid))
+   plot( as.numeric(names(tt)), tt, type="h", lwd=3, lend=1,
+         xaxs="i", xaxt="n", xlim=c(0,35),
+         yaxs="i", yaxt="n", ylim=c(0,2900) )
+   text( 10, 2000, vn, cex=1.0, font=2, adj=c(0,0) )
+   # Finally return number of persons with at least one valid value of
+   # the variable:
+   c( sum(tt), median( rep(as.numeric(names(tt)),tt) ) )
+ }
> par( mfrow=c(4,6), oma=c(4,4,1,1), mar=c(0,0,0,0), mgp=c(3,1,0)/1.6, bty="n" )
> for( i in 1:length(wh) )

```

```

+   {
+     nval[i,] <- n.vis( wh[i] )
+     if( i %in% (0:3*6+1) ) axis(side=2)
+     if( i > 16 ) axis(side=1)
+   }
> mtext( "No. of persons", side=2, line=2, outer=TRUE, las=0 )
> mtext( "No. of dates" , side=1, line=2, outer=TRUE )
> round( nval, 1 )

```

var	no. pers	median no. values
abdominalomfang	714	1
bmi	12993	1
blodglukose	591	1
middelblodglukoseepj	879	1
gluc	10943	12
hba1c	14810	15
diastoliskepj	4382	6
diastoliskhjemme	1987	1
diastolisk.arm	2749	2
systoliskepj	4382	6
systoliskhjemme	1987	1
systolisk.arm	2188	2
tchol	14618	4
hdl	14598	4
ldl	10976	3
vldl	13725	4
triglycerid	14586	4
cpeptid	11584	1
gfr	1406	4
egfr	8015	2
dualb	6274	6
ualbcrea	12744	5

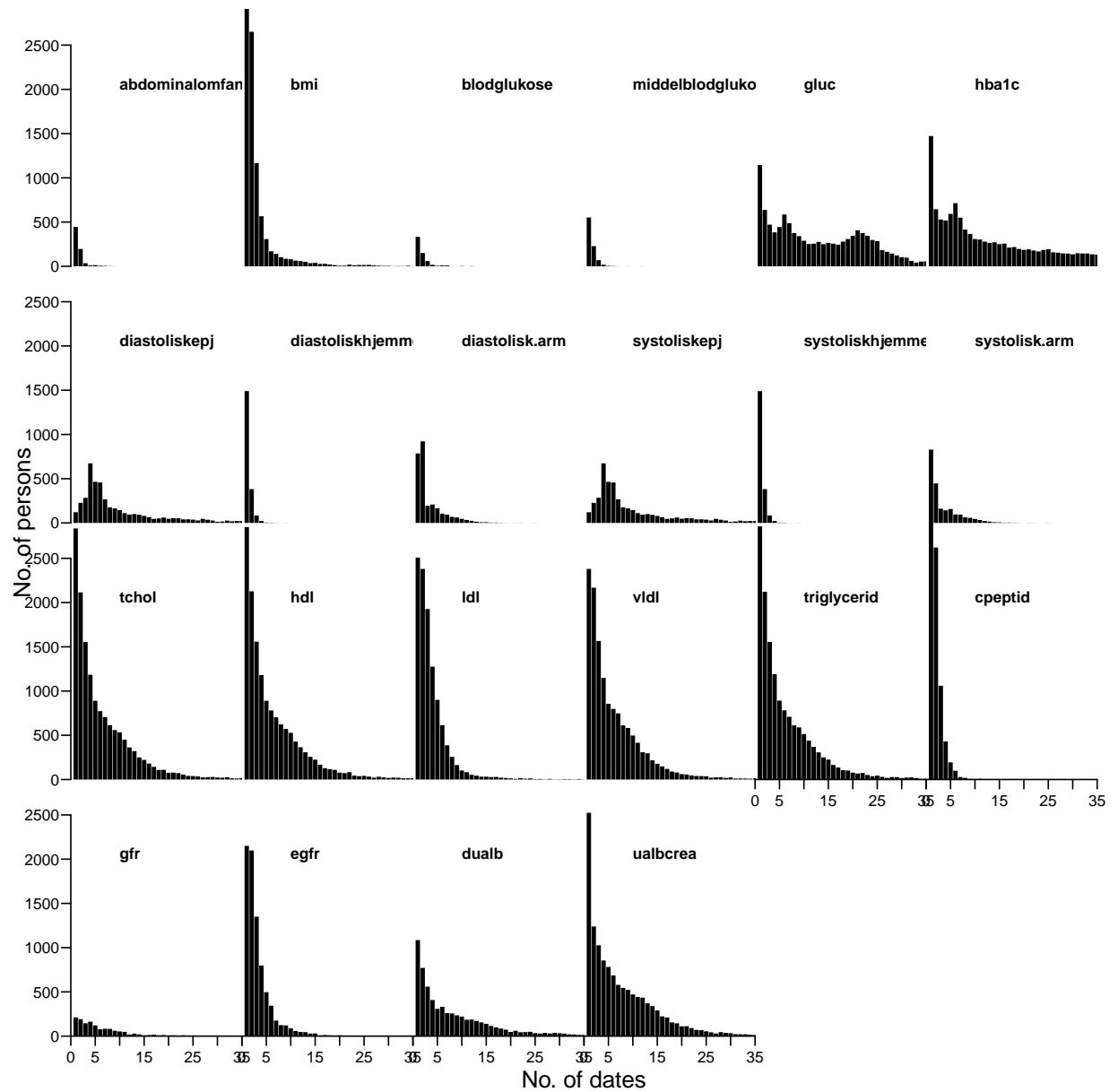


Figure 2.3: Number of recorded dates with valid values of specific measurements for patients in the extract. The total area (the sum of the bar heights) is the number of persons with at least one valid measurement of the variable.

# Chapter 3

## Analyses

```
> options( width=90 )
> library( Epi )
> load( file="./data/nef.Rda" )
```

### 3.1 Outcome data

Here is an overview of the perons with a valid date of ESRD:

```
> esr <- subset( nef, !is.na( doesrd ) )
> addmargins( tt <- with( esr, table(table(newid)) ) )
```

2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
2	2	3	6	9	6	7	3	9	9	11	17	8	9	7	11	12	6	2	9	6	6
24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
5	6	9	8	8	7	7	10	4	5	9	7	8	3	5	5	7	2	7	4	6	5
46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67
6	11	5	1	7	2	3	5	8	6	7	7	4	7	4	3	1	3	2	5	4	2
68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89
1	4	8	1	4	2	5	5	2	1	5	2	3	1	2	2	1	1	2	2	1	2
91	92	93	94	95	96	97	99	100	101	103	109	119	134	179	214	Sum					
1	1	1	2	2	1	1	2	1	1	1	3	1	1	1	1	478					

...so there are 478 persons with ESRD for analysis.

But we need to fish out all records with GFR measurements, and subsequently persons with at least two measurements of (e)GFR

```
> sdcR <- subset( nef, !is.na(egfr) & (dolab <= doesrd) )
> dim( sdcR )
[1] 6791 55
> 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
30	19	16	18	19	12	17	10	6	10	5	6	9	4	7	4	2	10	8	4	6	3
23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44
6	5	7	4	5	9	3	7	2	7	2	8	5	6	6	5	4	3	3	3	4	3
45	46	47	49	51	52	53	54	55	57	62	66	85	162	Sum							
1	2	2	5	2	1	1	3	3	2	1	1	1	1	358							

```
> # 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	23
19	16	18	19	12	17	10	6	10	5	6	9	4	7	4	2	10	8	4	6	3	6
24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
5	7	4	5	9	3	7	2	7	2	8	5	6	6	5	4	3	3	3	4	3	1
46	47	49	51	52	53	54	55	57	62	66	85	162	Sum								
2	2	5	2	1	1	3	3	2	1	1	1	1	328								

so after excluding persons with insufficient

Since we are going to analyse 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(-15,0,0.25) )
```

## 3.2 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 stage renal disease outcome (ESRD), and model how the trajectories of GFR are in these individuals *before* the endpoint ESRD. 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, !is.na(doesrd) )
> esrd$age <- esrd$dolab - esrd$dob
> with( esrd, round( quantile( ttESRD, 0:10/10 ), 1 ) )
  0%   10%   20%   30%   40%   50%   60%   70%   80%   90%  100%
-14.1 -9.7 -7.8 -6.4 -5.3 -4.2 -3.2 -2.4 -1.5 -0.7  0.0
> ( kn <- seq(-14,0,,5) )
[1] -14.0 -10.5 -7.0 -3.5  0.0
> MM <- Ns( esrd$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,] 0.3238895 -0.04807288 -0.1121177
[2,] 0.3299052 -0.01051803 -0.1310351
[3,] 0.3184025  0.05432337 -0.1502266
[4,] 0.2920431  0.11125574 -0.1573147
[5,] 0.2641092  0.15370524 -0.1578031
[6,] 0.2067223  0.21930847 -0.1507125
```

```
> sum( MM[,1]*esrd$ttESRD )
[1] -4.33328e-11
> sum( MM[,2]*esrd$ttESRD )
[1] 6.686903e-11
> sum( MM[,3]*esrd$ttESRD )
[1] 1.788074e-11
```

We have now set up data to fit the model; the columns **x1**, **x2** and **x3** 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.

Now we want to see how things pan out with either 3 or 4 classes and

```
> system.time(
+ spl3 <- hlme( lgFR ~ x1 + x2 + x3 + ttESRD + age + sex,
+             mixture = ~ x1 + x2 + x3 + ttESRD,
+             random = ~ ttESRD,
+             subject = 'newid',
+             ng = 3, nwg=FALSE,
+             # convB = 0.001, convG = 0.001, convL = 0.001, verbose=TRUE,
+             data = transform( esrd, lgFR=log(GFR) ) ) )
> system.time(
+ spl4 <- hlme( lgFR ~ x1 + x2 + x3 + ttESRD + age + sex,
+             mixture = ~ x1 + x2 + x3 + ttESRD,
+             random = ~ ttESRD,
+             subject = 'newid',
+             ng = 4, nwg=FALSE,
+             # convB = 0.001, convG = 0.001, convL = 0.001, verbose=TRUE,
+             data = transform( esrd, lgFR=log(GFR) ) ) )
> save( spl3, spl4, file="../data/splf.Rda" )

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