

Clinical nephropathy from SDC

SDC

July 2016

<http://bendixcarstensen.com/>

Version 5

Compiled Wednesday 8th February, 2017, 16:53
from: /home/bendix/sdc/proj/HKPWH/r/SDC.tex

Bendix Carstensen Steno Diabetes Center, Gentofte, Denmark
& Department of Biostatistics, University of Copenhagen
bxo@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

1	Reading SDC data	1
1.1	Reading the SDC clinical data	1
1.1.1	Utilities	1
1.1.2	The Stata dataset	1
1.2	Dates and events	3
1.2.1	Date problems	4
1.2.2	Overview of dates	13
1.2.3	Date variable relations	15
1.3	GFR and other renal measurements	15
1.3.1	Renal endpoints	17
1.4	Exporting data for comparative analyses	19
	References	22

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.

1.1.1 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)]
+ }
```

1.1.2 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/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 )

abdominalomfang      units
b12                  "cm"
blodglukose          "pmol/l"
bmi                  "mmol/l"
cpeptid              "kg/m^2"
diastoliskepj        "pmol/l"
diurese              "mm hg"
dualb                "ml"
egfr                 "mg/d"
gad                  "ml/min"
gfr                  "kiu/l"
haemoglobin          "ml/min"
hba1c                "mmol/l"
hdl                  "mmol/mol"

```

```

height          "m"
hvillepuls      "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	Genoptr\	xe6ning		NA	Denmark			NA	NA			NA	NA
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 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	dctype	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	diastoliskep	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	0				
16887		NA	NA	NA	NA	NA	NA	0				
	pcreatinin	region	rygning	systoliskep	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	dctype	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	diastoliskep	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	systoliskep	trans	triglycerid	tsh	ualbcrea	vldl	weight		
72865		NA	<NA>		NA	NA	NA	NA	NA	NA		
72866		NA	<NA>		NA	NA	NA	NA	NA	NA		
	dodm											
72865				1981								
72866				1981								
	newid	sex	dctype	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	diastoliskep	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	0				
104324		NA	NA	NA	NA	NA	NA	0				
	pcreatinin	region	rygning	systoliskep	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	dctype	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 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		
			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	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	


```

      user  system elapsed
284.308    0.117 284.597

> # 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
```

1.2.2 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    55
> 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

```

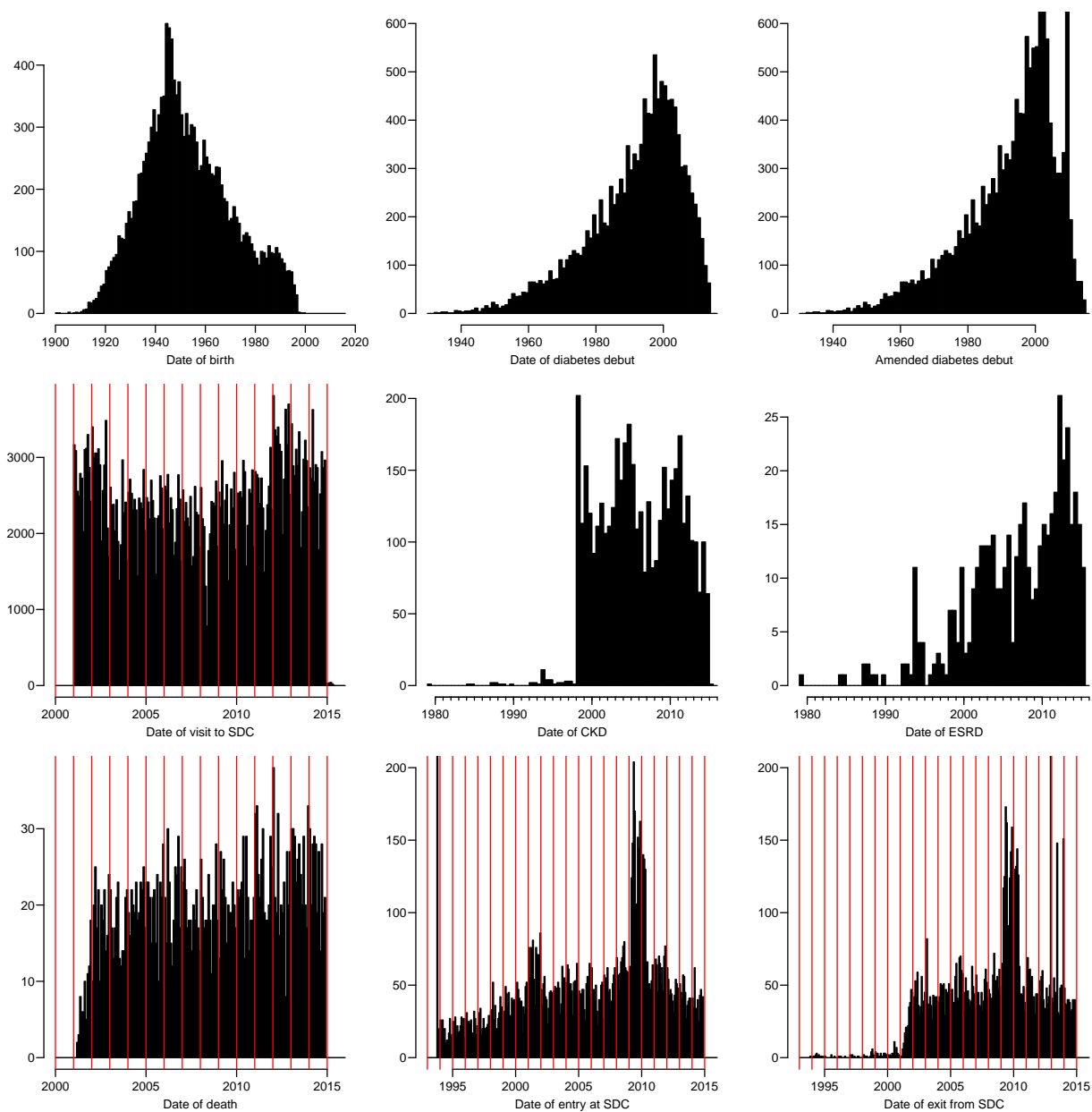


Figure 1.1: *Histograms of various dates from the dataset.*

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.

1.2.3 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")} )
```

1.3 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) ) ) )
```

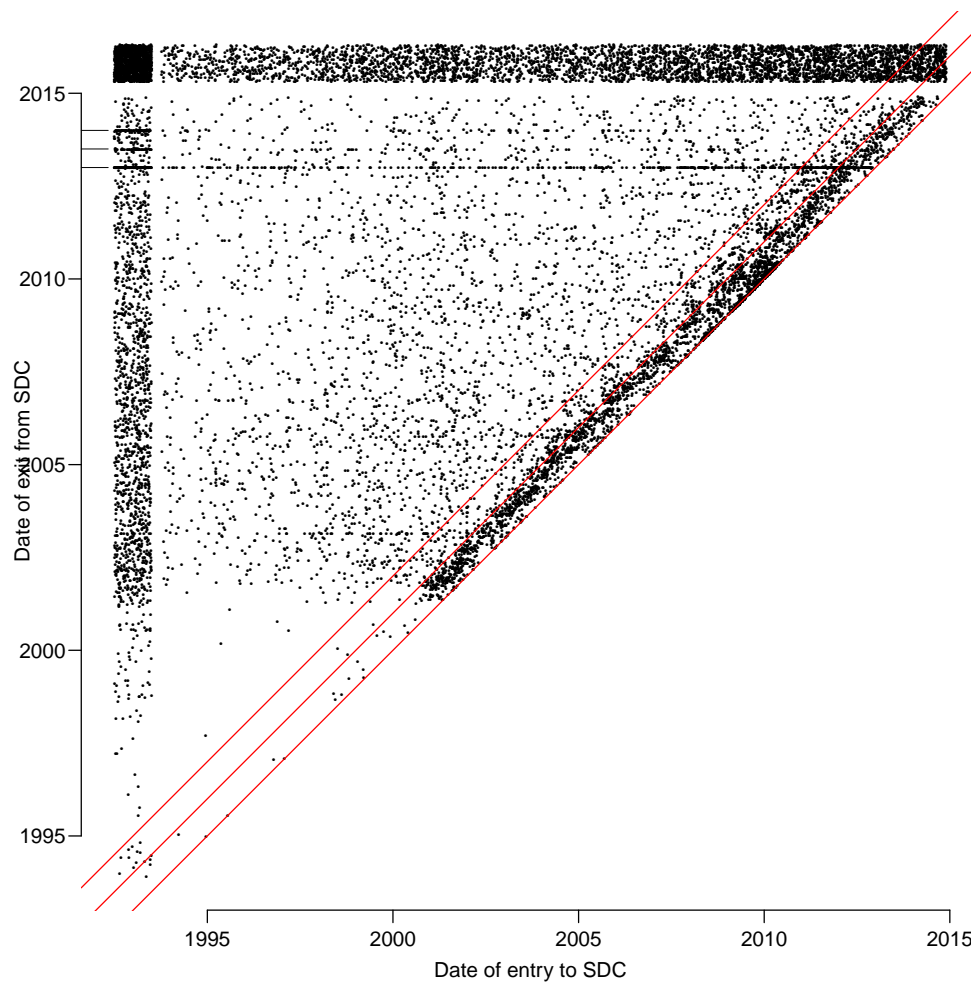


Figure 1.2: Joint distribution of entry and exit dates to SDC. The band to the left are those with date of entry coded as 1993-10-03, and the band at the top those with date of exit missing.

```

      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"

```

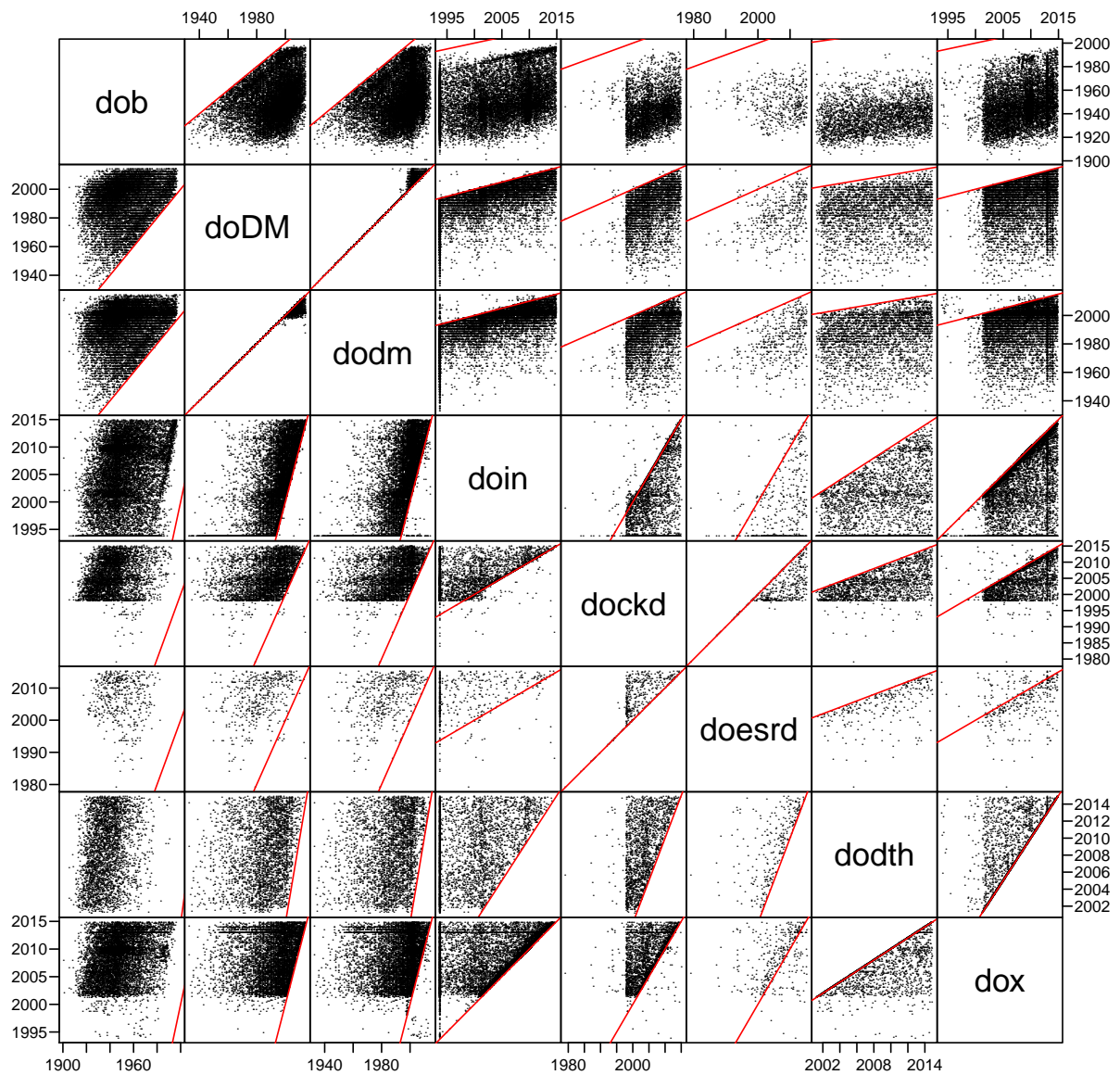


Figure 1.3: Date variables in the SDC clinical dataset. Each dot represents one person. The red lines are the identity lines, meaning that all points should be on the same side of the lines since the date variables are listed in approximately ascending order.

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

1.3.1 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)) )
```

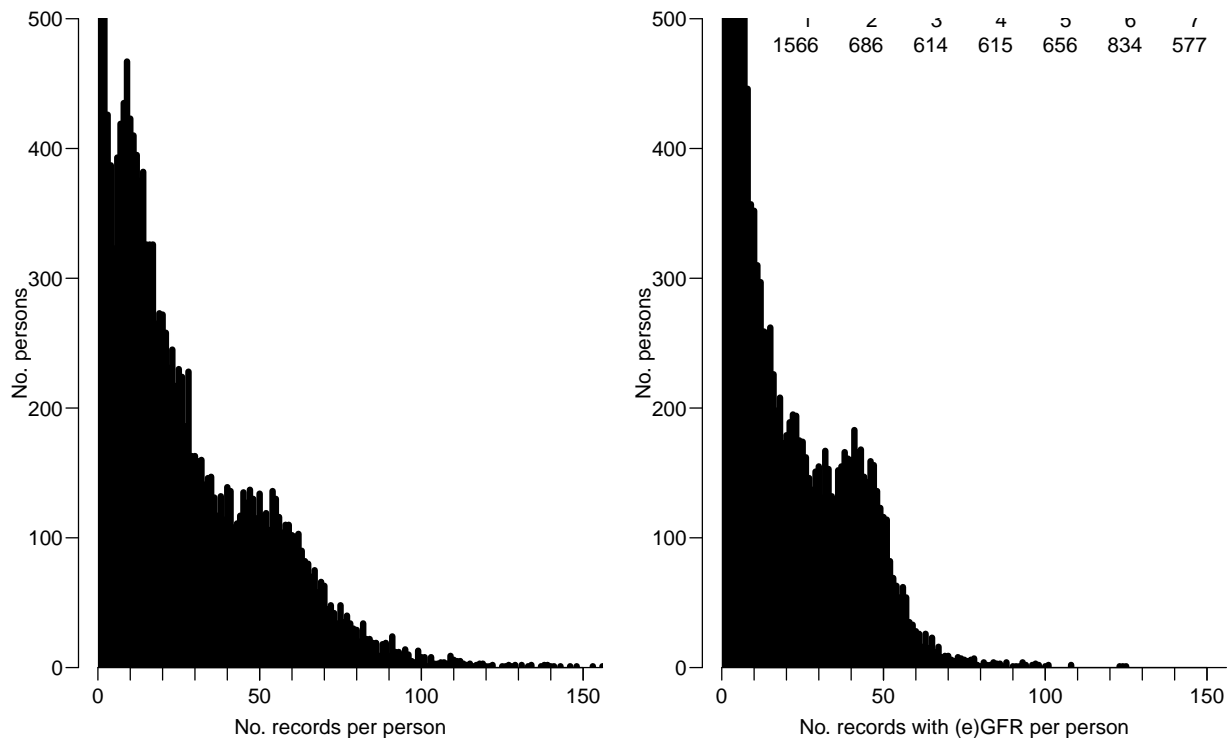


Figure 1.4: Persons in the database classified by the number of records in the dataset, resp. number of records with GFR or eGFR measurements.

```

      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=doesrld<=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(doesrld) )
> 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" )
```

1.4 Exporting data for comparative analyses

The following reads the original Danish nephropathy data and outputs it for comparable analyses with the PWH-HK data:

```

> library(Epi)
> load("../data/nef.Rda")
> lls()
  name mode class      size
1 nef  list data.frame 417363 55
> names( nef )
 [1] "newid"      "dolab"      "sex"        "dmtype"
 [5] "alkohol"    "civilstandskode" "motion"     "region"
 [9] "rygning"    "dob"        "doDM"       "age"
[13] "esrd"       "doesrd"     "ckd"        "dockd"
[17] "dth"        "dodth"      "doin"       "dox"
[21] "abdominalomfang" "b12"       "black"      "blodglukose"
[25] "bmi"        "cpeptid"    "diastoliskep" "diurese"
[29] "dualb"      "duplicates" "egfr"       "gad"
[33] "gfr"        "haemoglobin" "hba1c"      "hdl"
[37] "height"     "hvilepuls"  "ldl"        "middelblodglukos"
[41] "migrant"    "pcreatinin" "systoliskep" "trans"
[45] "triglycerid" "tsh"        "ualbcrea"   "vldl"
[49] "weight"     "dodm"       "prev"       "GFR"
[53] "ren.st"     "alb.st"     "ckd.st"

> wh <- c(1,3,10,11,4,19,20,18,2,31,33)
> whnam <- c("id","sex","dob","dodm","dmtype","doin","dox","dodth","dolab","egfr","gfr")
> data.frame( wh, names(nef)[wh], whnam )
  wh names.nef..wh. whnam
1  1      newid      id
2  3         sex      sex
3 10         dob      dob
4 11        doDM     dodm
5  4        dmtype  dmtype
6 19         doin     doin
7 20         dox      dox
8 18        dodth    dodth
9  2         dolab    dolab
10 31         egfr     egfr
11 33         gfr      gfr
> DKnef <- nef[,wh]
> table( nef$dmtype, nef$sex )
           Male Female
andre former    3822   1920
ikke spec.      770   1378
type 1          112160 106119
type 2          108807  74986
type ikke angivet  4311   3017
> DKnef$dmtype <- Relevel( factor( nef$dmtype ),
+                           list( T1D = 4,
+                               T2D = 5,
+                               Oth = 2,
+                               Unkn = c(1,3,6) ) )
> table( nef$dmtype, DKnef$dmtype )
           T1D    T2D    Oth    Unkn
andre former    0     0     0     73
ikke spec.      0     0     0    2148
type 1          218279  0     0     0
type 2           0 183793  0     0
type ikke angivet  0     0     0   7328

```

```
> names(DKnef) <- whnam
> head( DKnef )
```

```
      id sex      dob dodm dmtyp      doin dox dodth      dolab      egfr gfr NA
1 10000 Male 1956.201 2000 type 2 2009.937 NaN   NaN 2001.828      NaN NaN T2D
2 10000 Male 1956.201 2000 type 2 2009.937 NaN   NaN 2009.937 100.14849 NaN T2D
3 10000 Male 1956.201 2000 type 2 2009.937 NaN   NaN 2010.052  96.59035 NaN T2D
4 10000 Male 1956.201 2000 type 2 2009.937 NaN   NaN 2010.263  98.27923 NaN T2D
5 10000 Male 1956.201 2000 type 2 2009.937 NaN   NaN 2010.320  98.77580 NaN T2D
6 10000 Male 1956.201 2000 type 2 2009.937 NaN   NaN 2010.583  96.23064 NaN T2D
```

```
> str( DKnef )
```

```
'data.frame':      417363 obs. of  12 variables:
 $ id   : int  10000 10000 10000 10000 10000 10000 10000 10000 10000 10000 10000 ...
 $ sex  : Factor w/ 2 levels "Male","Female": 1 1 1 1 1 1 1 1 1 1 1 ...
 $ dob  : num  1956 1956 1956 1956 1956 ...
 $ dodm : num  2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 ...
 $ dmtyp: chr   "type 2" "type 2" "type 2" "type 2" ...
 $ doin : num  2010 2010 2010 2010 2010 ...
 $ dox  : num   NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN ...
 $ dodth: num   NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN ...
 $ dolab: num  2002 2010 2010 2010 2010 ...
 $ egfr : num   NaN 100.1 96.6 98.3 98.8 ...
 $ gfr  : num   NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN ...
 $ NA   : Factor w/ 4 levels "T1D","T2D","Oth",...: 2 2 2 2 2 2 2 2 2 2 ...
```

```
> summary( DKnef )
```

id	sex	dob	dodm	dmtyp	doi
Min. : 1	Male :229918	Min. :1901	Min. :1933	Length:417363	Min. :
1st Qu.: 3868	Female:187445	1st Qu.:1941	1st Qu.:1980	Class :character	1st Qu.:
Median : 7705		Median :1951	Median :1991	Mode :character	Median :
Mean : 7686		Mean :1953	Mean :1989		Mean :
3rd Qu.:11557		3rd Qu.:1964	3rd Qu.:1999		3rd Qu.:
Max. :15327		Max. :2000	Max. :2014		Max. :
			NA's :15636		NA's :
dox	dodth	dolab	egfr	gfr	NA
Min. :1994	Min. :2001	Min. :2001	Min. : 2.89	Min. : 5.0	T1D :21
1st Qu.:2007	1st Qu.:2007	1st Qu.:2004	1st Qu.: 86.08	1st Qu.: 46.0	T2D :18
Median :2011	Median :2010	Median :2008	Median : 94.49	Median : 69.0	Oth :
Mean :2010	Mean :2010	Mean :2008	Mean : 93.57	Mean : 70.8	Unkn:
3rd Qu.:2013	3rd Qu.:2013	3rd Qu.:2012	3rd Qu.:105.23	3rd Qu.: 94.0	
Max. :2015	Max. :2015	Max. :2015	Max. :216.98	Max. :175.0	
NA's :236181	NA's :340801		NA's :137895	NA's :412129	

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

Bibliography