DCRC: Tabulation outline

Diabetes and Cancer Research Consortium
Version 4

Compiled Tuesday 3rd January, 2012, 13:14
from: `C:/Bendix/Steno/DM-register/NDR/projects/Cancer/Consortium/papers/tabulate/tabulate.tex`

Bendix Carstensen
Steno Diabetes Center, Gentofte, Denmark
& Department of Biostatistics, University of Copenhagen
bxc@steno.dk
http://www.biostat.ku.dk/~bxc/

Contents

Chapter 1

Introduction

This document is meant as a guideline for a general tabulation of follow-up data for the contributing members for the Diabetes and Cancer Research Consortium. These are a set of desirable criteria to meet; the purpose is to establish summary datasets (tables) that allows joint analysis of cancer incidence / mortality across centers, with the specific aim of evaluating the effects of exposure to certain drugs used in diabetes treatment, that be causal or assignment effects.

Chapter 2

Exposures (covariates)

2.1 Drugs

The following pharmaceutical exposures are of interest:

- Metformin (A10BA)
- SU (A10BB))
- TZDs (A10BG)
- Insulin (A10A) — as well as a subdivision in different analogs

This means that follow-up should be classified by indicator variables of whether these drugs are actually being taken in any given follow-up interval. Preferably the dosage of these should also be coded separately for each interval.

2.2 Timescales

The following timescales are of interest:

- Time since AD 0 (current calendar time)
- Time since birth (current age)
- Time since DM diagnosis (current disease duration)
- Time since first Metformin dispensation
- Time since first SU dispensation
- Time since first TZD dispensation
- Time since first Insulin dispensation

For the last 5 timescales, persons who are not (yet) diagnosed with diabetes should be coded 0; persons diagnosed with diabetes and/or on any of the drugs but where duration is unknown should be coded NA (missing, Not Available).

Insisting on including all these timescales will inevitably limit data, since the dates of start will not be known for some (i.e many) patients, who therefore will be excluded.

The latter will enable analyses investigating duration effects excluding persons without duration information, as well as analyses including all persons ignoring the duration variable.

Also note that these duration variables are defined as time since first dispensation, that is they keep increasing, even if a given drug is stopped. Moreover, we will not have any handle on time off a drug, but we will know by the interaction between (time since first dispensation > 0) and the indicator of the drug being taken whether persons off the drug have a higher or lower risk that those who have never been on it.

We require that age and calendar time at follow-up as well as date of birth be tabulated in fairly small intervals — emerging evidence suggests that there may be quite dramatic changes of cancer occurrence in the first few months after diagnosis of diabetes.

If we require that the 5 duration timescales be tabulated in 6-month intervals, we may very well produce a very large dataset indeed, most likely substantially exceeding 10,000,000 records (see appendix).

2.3 Timescale practicalities

Allocation of events and follow-up time to intervals on the many timescales formally requires that the dataset be split on these.

Technically this can be done in Stata using the command `stsplit`. In SAS by using the macro `%Lexis`, available as `http://staff.pubhealth.ku.dk/~bxc/Lexis/Lexis.sas`, which contains guidance to the use in the file itself. In R is a machinery called `splitLexis`, but since R keep all data in memory this may be prohibitive by the sheer data size on a 32bit machine.

However, it would be a better approach to split follow-up only on one time scale and compute the values of all other timescales at the beginning of each interval. The advantage of the `Lexis` machinery in R is that this is automatically done for any time-scale defined.

However it is necessary to cut the follow-up for any patient initiating a particular drug at the time of initiation. This should be done before making the split of the follow-up on, say, age.

2.3.1 Covariates

The indicators and the timescales define 11 variables, so including sex and date of birth we will have 13 explanatory variables.

2.4 Outcomes

We propose to include follow-up only until first primary cancer, and censor persons at this event. This will simplify analyses, since the follow-up time will be the same for all cancer incidence outcomes considered.

In the cancer epidemiology literature there are varying practices on this point; some prefer to follow persons till the occurrence of the primary cancer of interest, disregarding earlier occurrence of other primary cancers. By the same token, only persons with the particular cancer diagnosed prior to start of follow up should be excluded.

The following outcomes should be tabulated for all:

- Follow-up time (person-years) before death.
- Death.
- Follow-up time (person-years) before first primary cancer of any kind.
- Any primary cancer.
- Any primary cancer except non-melanoma skin cancer.
- (all the other cancers — specify; ICD10 codes.)

These outcomes thus define $5 + \{\text{number of cancer sites}\}$ variables in the dataset.

2.5 Non-DM follow-up

Not all studies have access to a full database of the entire population, but only to demographic data from the statistical bureau (population size and hence derived population follow-up time), and to cancers for the entire population, typically by sex, age, calendar time and date of birth.

In this case, the follow-up time and cancer cases in the DM population must be subtracted from that of the total population to give the cases and follow-up time in the non-DM population.

2.6 Comparison groups

2.7 Specifics

A number of specific features of the different studies must be taken into account:

Scotland The diabetes classification is incomplete prior to 2003(????) and hence the follow-up among those coded as non-diabetics contains a fraction of DM patients. Thus analyses that compares rates between non-DM and DM are not valid for the period prior to this date. But comparisons *internally* in the group of DM patients are.

Canada (BC) The data is not a complete enumeration of the follow-up in the population, but only among diabetes patients and a sample of non-DM persons matched to the DM-persons at date of diagnosis. Hence, the DM patients can only be followed from the date of DM, and the matched non-DM persons only from the matching date.

THIN/GPRD The cancer diagnoses are based on extracts from GP databases, and are therefore less reliable, and particularly some persons with a previous cancer may be included.

Chapter 3

Design

3.1 Follow-up

The advantage of a follow-up study is the flexibility in definition of time-dependent exposures, that be medication exposure or clinical features measured.

3.1.1 Complete population registers

This is the most comprehensive and most flexible, since the databases can be used for all kinds of analyses, including estimation of absolute rates of cancer and hence also cumulative probabilities of cancer occurrence.

3.1.2 Matching from complete population registers

As an alternative to follow-up of the entire population, one may choose a random sample of the background population. In principle any sample of the population could be used as long as the selection of the population sample is independent of 1) the outcome of interest (in this case cancer) and 2) the exposure of interest (in this case diabetes and medication).

A total random sample of the population would presumably be a waste of money (if a per record cost is in force), since both cancer and diabetes occur in older ages. As the exposure of primary interest is diabetes, one option would be to select non-diabetes persons matched to diabetes cases. That is for each new case of diabetes, select one or more persons from the population without diabetes (and cancer) at the point of diagnosis of the diabetes case. Point of diagnosis can be defined on any number of variables, but presumably sex, current age and current calendar time (and hence date of birth) would mostly be used. If socioeconomic variables were to be included, these could be used for matching too.

Note that the population sample can include persons that later acquire diabetes, and the follow-up of these must be in the non-diabetes group until their date of diagnosis of diabetes, and from then on in the diabetes group. They will presumably be in the diabetes group already, and hence their follow-up will in practical terms just cease at the date of diabetes. If these persons were excluded from follow-up, we would selectively exclude follow-up among persons known to develop diabetes, that is persons with higher risk of diabetes. To the extent that these risk factors are also risk factors for cancer, we would potentially exclude more cancer cases, than would be the case if they were included.

The analysis of a matched study like this would be exactly as for a complete register, one would have to include the matching variables in the analysis. Note that the point of matching for the population sample has no meaning — it is just a random data in these persons' lives, so there is no such variable as time since inclusion for the population sample.

The disadvantage in relation to a total population sample is only the extra work in selecting the matching sample, and the slight disadvantage of the smaller number of cancer cases in the population comparison group. The latter is likely not of any big importance, since the factor limiting the precision in the comparisons is the number of cancer cases in the diabetes group. For the rarer cancer forms it might however be a disadvantage that the population sample could be so small that that the modeling of the population rates becomes unstable.

3.1.3 Ad-hoc cohorts

3.2 Case-control

Case-control studies has the advantage of simplicity of analysis. Even if large population registers are available it can provide analytical advantages to sample all cases of cancer and match them to a sample of the persons, who are free of the cancer at the time of the case. Note that the persons who get the cancer diagnosis later can be included as controls as well at any point in time (age) before they are diagnosed. Also not that in this type of studies the matching time is crucial, because the covariates (notably drug exposure) must be computed at the matching point in time.

3.3 Reporting

The group will follow the STROBE (STrengthening the Reporting of OBservational studies in Epidemiology) guidelines for reporting observational studies, see `http://www.strobe-statement.org/`.

Chapter 4

The tabulation squeeze

In pharmacoepidemological studies of diabetes we are interested in many timescales beyond the fundamental three, current age, current calender time and disease duration. Each drug of interest will typically require two timescales, namely time since initiation of the drug and time since the cessation of it.

The point of this note is not to discuss the finer points of the definition of these variables, here we shall just assume that algorithms are available to define all relevant time scale variables at any desired point of follow up for all persons in the study.

The purpose of this note is to discuss practical data processing problems with many timescales.

4.1 Data requirement for follow-up data

If multiple timescales are to be accommodated, it is required that the follow up time is subdivided by each of these. Splitting of follow-up time by age and calendar time as well as by a number of time scales will result in a very large number of units from each patient, and potentially also a very large number of cells in the required cross-classification of timescales.

4.2 The tabulation squeeze

If four drugs and diabetes are to be classified by duration in say 6-month intervals, then we will with 15 years of follow up have 30 intervals on each time scale, that is potentially $30^5 = 24.3$mio. intervals, which additionally must be classified by age, calendar time and diabetes duration. A substantial fraction of these potential combinations will of course be empty, but with the additional tabulation by age and period, we can easily run into hundreds of millions of combinations, which currently is not feasible as analysis unit.

4.3 A practical solution

It should be noted that it is only the diabetes patients' follow-up that need subdivision by drug-exposures. And so far we have only a few hundred thousand patients in each data base. So if the follow-up of diabetes patients is only split by time since diagnosis (duration of diabetes), in say 6-month intervals, we will have up to 30 intervals per person. In the Danish diabetes and cancer study there is about 1,000,000 person-years among DM patients, so this tabulation would result in some 2,000,000 intervals, which is in the range of analytical possibilities.

The advantage of this is that we can define all of the required timescales for any of these intervals, so this approach is robust to inclusion of any number of time scales, as the number of units will stay the same regardless of further time-scales being added.

The follow-up of the non-diabetic population is still classified and tabulated by current age, calendar time and date of birth. In order to make the follow-up of the diabetes patients comparable to this, the age and calendar time assigned to each interval should formally be the age and calendar time 3 months (*i.e.* half the tabulation length) after the left endpoint of the interval (which for most, but not all, intervals will the midpoint). This is because we then have the incidence rate which is assumed constant in each interval allocated to the correct point on the timescale. However, if analysis is preformed by using variables derived from the timescales age and calendar time, it is preferable to use the values at the beginning of the intervals, because we then preserve the relationship between the timescales within each person.

Likewise, the duration and cumulative exposure variables, should correspond to the left endpoint of the intervals, because we otherwise would be conditioning on the future.

This way we will be able to produce a dataset which in the case of Danish data will have some 2 million records, and which can accommodate any number of timescales for analysis. Hence, it will only be the the number of events that limits the complexity of the models, not the tabulation possibilities. If we instead of 6-month intervals use 2-month intervals (which would presumably be the smallest possible, due to the limitations in the precision of recording of dates in the two registers), we would have about 50 records per person, so some 5–6 million records in total.

4.4 Confidentiality

The resulting dataset will be a dataset which has very large numbers of person-years for each age, period, cohort class for the non-diabetic population and very small amounts of follow-up for each combination of the many timescales for the diabetes population. Some (presumably most) contributions from the diabetes population will only contain follow-up data from one person.

It will however be totally uninformative about the the persons identity, because it will only concern a small piece of the follow-up from the person, and there will be no way to link this piece of information to the rest of the information from the same person. Hence, there will be no way to link any of the follow-up tabulated this way back to the individuals. Unless, of course, all the information contained in the record is known from some other source, in which case the confidentiality issue would be somewhere else.

Chapter 5

Construction of covariates

This chapter is a *very* technical explanation of how to construct the relevant cumulative exposure variables using R. It contains an exposition of the considerations that are behind construction of variables, leading to the construction of an R-function that does the job for drug purchase records where the dose intensity (daily dose) may or may not be recorded.

5.1 Translating dose / amount into exposure covariates

Suppose records of drug purchase are available, and that the amount is available at each purchase too.

If an assumed dose rate (dose per time) is known for each drug purchase, then the purchased amount divided by the dose rate equals the period of drug coverage. Thus, each purchase record can be transformed into a period covered, namely from date of purchase (dop) to date of purchase plus the length of the period covered, that is code of this kind (amt–amount, dpt–dose per time):

```
> drug.start <- dop
> drug.end   <- dop + amt/dpt
```

Note that we with this sort of calculation assume that medication is consumed at a constant rate (dpt). But this could easily be a prescription-specific figure.

5.1.1 Insulin

A special case is insulin, where there is very rarely any indication of the dosage. Hence we basically have no handle on the period covered by each prescription taken out. In this case we need a machinery that basically assumes that each purchase is continuously consumed over the period till the next purchase. This is handled in a separate section.

5.2 Variables / scales of interest

Drug exposures vary by time, so whether a cohort or a case-control approach is used we need a machinery that defines exposures at any point during follow-up.

In the assessment of disease risk as a function of drug exposure the following variables may be of immediate interest for each drug:

- `tfi`: time from initiation, *i.e.* time since first use of the drug (in practice time since date of first purchase).
- `tfc`: time from latest cessation, that is the time since the end of the coverage period from the latest purchase.
- `cdur`: cumulative time on the drug.
- `cdos`: cumulative dose of the drug; in principle this amounts to the amount purchased, but using the assumption of constant consumption rate for each purchase, we can compute it at any given date between purchases too..
- `ldos`: lagged cumulative dose, that is the cumulative dose as it was a given time ago.

Note that the first three variables mentioned are measured in (calendar) time while the latter two are measured in dose units. Thus when using them in models as linear terms the coefficients will have different units.

However, we will not use them as linear terms, but in a non-linear form. Based on previous experience we will expect that for practically any disease outcome there will be an excess in the first period after initiation of a drug, and possibly also in the first short period after cessation of a drug.

The cumulative time on the drug and the cumulative dose of the drug will be very closely correlated, so it will in practice be difficult to accommodate both in a model.

Also noteh that for all the variables mentioned here, it is assumed that the entire medication history is kown. If this is not the case, the only variables that can be meaningfully defined are “current dose”, “currently on the drug” but possibly not even “ever on the drug”.

5.3 Implementation in R

There are basically two different scenarios for calculating of the cumulative dose; one that uses available dosage information, and one that ignores this.

To show how these variables are constructed we create a bogus dataset for illustration and develop the function on that:

```
> # Construct a dataset of medication records for three persons
> n <- c( 10, 17, 8 )
> dop <- c( 1995.2+cumsum(sample(1:4/10,n[1],replace=TRUE)),
+           1996.7+cumsum(sample(1:4/10,n[2],replace=TRUE)),
+           1998.1+cumsum(sample(1:4/10,n[3],replace=TRUE)) )
> amt <- sample( 1:2/10, sum(n), replace=TRUE )
> dpt <- sample( 6:8/10, sum(n), replace=TRUE )
> PUR <- data.frame( id = rep(1:3,n),
+                    dop = dop,
+                    amt = amt,
+                    dpt = dpt )
> round( PUR, 3 )
```

```
   id    dop amt dpt
1   1 1995.3 0.1 0.6
2   1 1995.5 0.1 0.7
3   1 1995.9 0.2 0.8
4   1 1996.0 0.1 0.8
5   1 1996.1 0.2 0.8
6   1 1996.5 0.2 0.8
7   1 1996.8 0.2 0.6
8   1 1996.9 0.1 0.6
9   1 1997.3 0.1 0.6
10  1 1997.5 0.2 0.7
11  2 1997.1 0.1 0.6
12  2 1997.5 0.1 0.8
13  2 1997.6 0.1 0.7
14  2 1998.0 0.2 0.7
15  2 1998.2 0.2 0.7
16  2 1998.4 0.2 0.8
17  2 1998.8 0.2 0.7
18  2 1999.2 0.2 0.6
19  2 1999.3 0.1 0.6
20  2 1999.6 0.1 0.6
21  2 2000.0 0.2 0.7
22  2 2000.2 0.1 0.8
23  2 2000.6 0.1 0.6
24  2 2000.8 0.2 0.8
25  2 2000.9 0.1 0.8
26  2 2001.3 0.1 0.8
27  2 2001.5 0.2 0.8
28  3 1998.3 0.2 0.6
29  3 1998.6 0.2 0.8
30  3 1998.9 0.2 0.7
31  3 1999.0 0.1 0.7
32  3 1999.4 0.1 0.8
33  3 1999.7 0.1 0.6
34  3 1999.8 0.2 0.7
35  3 2000.0 0.2 0.6
```

We also need to construct a simple data frame for follow-up periods for these 3 persons:

```
> fu  <- data.frame( id = 1:3,
+                   doe = c(1995,1997,1996)-3:1/4,
+                   dox = c(2001,2003,2002)+1:3/5 )
> round( fu, 2 )
```

```
  id     doe    dox
1  1 1994.25 2001.2
2  2 1996.50 2003.4
3  3 1995.75 2002.6
```

So these two bogus datasets have the structure of input datasets from a prescription database and a database of follow-up of persons. Note that we are so far not concerned about the disease outcome, this paper only focuses on the meaningful construction of covariates at different times of follow-up.

In the first instance we will construct a dataset which for each person at a set of date have the cumulative dose at this date. Assuming linear increase in cumulative dose between these points will enable us to compute the cumulative dose at *any* of date.

5.3.1 The case with dose information

When dose per day is recorded, then we can use this to compute the exposed time associated with each purchase.

The dates of exposure for a particular purchase should be pushed so that the exposure start, `exp.start` say, is after the expiry of the coverage of the previous purchase, but never earlier than the end of the previous drug-coverage period.

We have also built in a facility to limit how far into the future a purchase can be pushed as exposure, via the `push.max` argument.

```
> use.amt.dpt <-
+ function( purchase,
+           push.max = Inf,
+             breaks,
+               lags = NULL,
+            lag.dec = 1 )
+ {
+ do.call( "rbind",
+ lapply( split( purchase, purchase$id ),
+         function(set)
+         {
+         np <- nrow(set)
+         if( np==1 ) return( NULL )
+         set <- set[order(set$dop),]
+         # Compute length of exposure periods
+         drug.dur  <- set$amt / set$dpt
+         # Put the exposed period head to foot
+         new.start <- min( set$dop ) + c(0,cumsum(drug.dur[-np]))
+         # Move them out so that the start of a period is never earlier than
+         # the dop
+         exp.start <- new.start + cummax( pmax(set$dop-new.start,0) )
+         # Compute the pushes
+         push.one <- exp.start - set$dop
+         # Revise them to the maximally acceptable
+         push.adj <- pmin( push.one, push.max )
+         # Revise the starting dates of exposure
+         exp.start <- exp.start - push.one + push.adj
+         # Revise the durations to be at most equal to differences between the
+         # revised starting dates
+         drug.dur  <- pmin( drug.dur, c(diff(exp.start),Inf) )
+         # Compute the end of the intervals
+         exp.end   <- exp.start + drug.dur
+         # Intervals in the middle not covered by the drug exposures - note
+         # also that we make a record for the last follow-date
+         followed.by.gap <- c( exp.start[-1]-exp.end[-length(exp.end)] > 0, TRUE )
+         # To facilitate
+         dfR <- rbind( data.frame( id = set$id[1],
+                                  dof = exp.start,
+                                  dpt = set$dpt ),
+                       data.frame( id = set$id[1],
+                                  dof = exp.end[followed.by.gap],
+                                  dpt = 0 ) )
+         dfR <- dfR[order(dfR$dof),]
+         # We now compute the cumulative dose at the end of the interval using
+         # interval length and dpt:
+         dfR$cum.amt <- with( dfR, cumsum( c(0, diff(dof)*dpt[-length(dpt)]) ) )
+         return( dfR )
+         } ) )
+ }
```

We can use this function to illustrate how the original purchases and the adjusted look for the three patients in the sample: First we compute the naïve coverage intervals from date of purchase and a period corresponding to the dose divided by the prescribed dosage.

```
> exp.start <- PUR$dop
> exp.end   <- with( PUR, dop+amt/dpt )
```

We then for illustration compute the coverage periods using two different setups, one where all purchases are assumed consumed, and one where some is considered lost:

```
> zz <- use.amt.dpt( PUR )
> zl <- use.amt.dpt( PUR, push.max=0.3 )
> zz$dur <- c(diff(zz$dof),NA)
> zl$dur <- c(diff(zl$dof),NA)
```

We can now plot the purchases, and the two different ways of converting these into coverage periods:

```
> par( mar=c(3,1,1,1), mgp=c(3,1,0)/1.6 )
> plot( NA, ylim=0:1, xlim=floor(range(zz$dof))+0:1,
+           bty="n", yaxt="n", ylab="", xlab="Date of follow-up")
> nr <- nrow( PUR )
> ys <- (1:nr-2+2*as.integer(PUR$id))/(nr+2)
> ym <- ave( ys, PUR$id, FUN=min )
> segments( exp.start   , ys-0.2/nr,
+           exp.end     , ys-0.2/nr, col="blue", lwd=2, lend=1 )
> with( subset(zz,dpt>0),
+       segments( dof         , ys-0.4/nr,
+                 dof + dur   , ys-0.4/nr, col="red", lwd=2, lend=1 ) )
> with( subset(zz,dpt>0),
+       segments( dof         , ym-0.4/nr,
+                 dof + dur   , ym-0.4/nr, col="red", lwd=2, lend=1 ) )
> with( subset(zl,dpt>0),
+       segments( dof         , ys-0.6/nr,
+                 dof + dur   , ys-0.6/nr, col="forestgreen", lwd=2, lend=1 ) )
> with( subset(zl,dpt>0),
+       segments( dof         , ym-0.6/nr,
+                 dof + dur   , ym-0.6/nr, col="forestgreen", lwd=2, lend=1 ) )
```

From figure 5.1 we see that there can be gaps in the exposure. These gaps are accounted for in the result from the function, they are given a dpt of 0.

5.3.2 The case without dosage information

In some circumstances there is no information in prescribed dose, that is if the dpt variable is not available, we must instead resort to computation of the cumulative dose at a given point in time as derived from the purchased amounts and the timings between these as illustrated in figure 5.2

```
> par(mar=c(6,6,1,1)/2, mgp=c(3,1,0)/1.6 )
> ptimes <- subset(PUR,id==1)$dop
> amount <- subset(PUR,id==1)$amt
> cumamt <- c(0,cumsum(amount))
>     np <- length(ptimes)
> ptimes <- c(ptimes,
+             ptimes[np]+
+             amount[np]/cumamt[np]*diff(range(ptimes)))
```

```
> plot( ptimes, cumamt, type="S", pch=16, col="red", lwd=1,
+       ylab="Cumulative purchase", xlab="Date of follow-up", bty="n",
+       xlim=range(ptimes), ylim=c(0,max(cumamt)) )
> segments( ptimes[-(np+1)], cumamt[-(np+1)],
+           ptimes[-(np+1)], cumamt[-(np+1)]+amount, lwd=4, col="red" )
> points( ptimes[-np-1], cumamt[-np-1], pch=16, col="blue", cex=1.5 )
> lines( ptimes, cumamt, lwd=4, col="blue" )
> lines( ptimes[c(1,np)],cumamt[c(1,np)],lty="13",lwd=2,col="blue")
```

The code to evaluate the cumulative dose at prespecified times uses the same example. As seen from figure 5.2 we need to add one more point to the vector of purchase dates, namely the point derived from an assumed consumption rate of the last purchase.

The exercise is to use the points (date of purchase, cumulative dose) (the blobs in figure 5.2) to predict the last point on the blue line, using the average slope over the exposure period (as indicated in the figure by the dashed line).

First we compute the cumulative amount prior to last purchase and the last purchased amount. Then the timespan from first to last purchase and then compute the average dose per time as the ratio of these two.

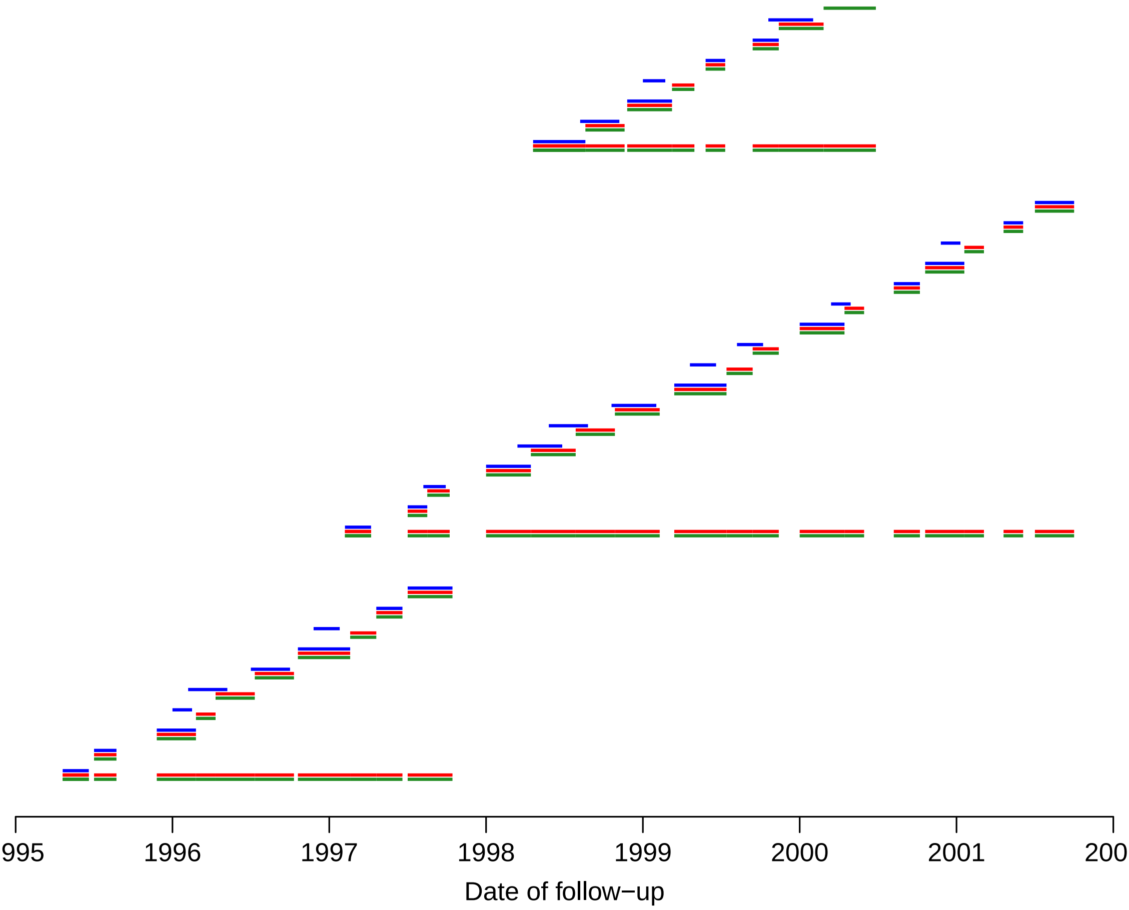

Figure 5.1: *Recorded, stacked exposure periods and corrected exposure periods assuming a maximal push of 0.3 years. At the bottom of each person is shown the exposure periods in total.*

This calculation should be done for each person separately. To this end we again use the function `split` which splits the data frame into a list of data frames, and we can then `lapply` to do what is needed:

```
> use.only.amt <-
+ function( purchase,
+           pred.win = Inf,
+             breaks,
+               lags = NULL,
+            lag.dec = 1 )
+ {
+ # Compute the cumulative dose at all purcase dates and at the last
+ # (unknown) future expiry date, computed based on previous
+ # consumption.  The resulting data frame has one more lines per person
+ # than no. of purchases.
+ do.call( "rbind",
+ lapply( split( purchase, purchase$id ),
+         function(set)
+         {
+         np <- nrow(set)
+         if( np==1 ) return( NULL )
+         set <- set[order(set$dop),]
+         # The points to include in the calculation:
+         # All dates after pred.win before last purchase,
+         # but at least the last two purchase dates,
+         wp <- ( set$dop > pmin( max(set$dop)-pred.win,
+                                 sort(set$dop,decreasing=TRUE)[2] ) )
+         # Cumulative amount consumed at each dop
+         cum.amt <- cumsum(c(0,set$amt))
+         # Average slope to use to project the duration last purchase
+         avg.slp <- diff(range(cum.amt[c(wp,FALSE)]))/
+                    diff(range(set$dop[wp]))
+         # Purchase dates and the date of last consumption
+         dof <- c( set$dop, set$dop[np]+set$amt[np]/avg.slp )
+         return( data.frame( id = set$id[1],
```

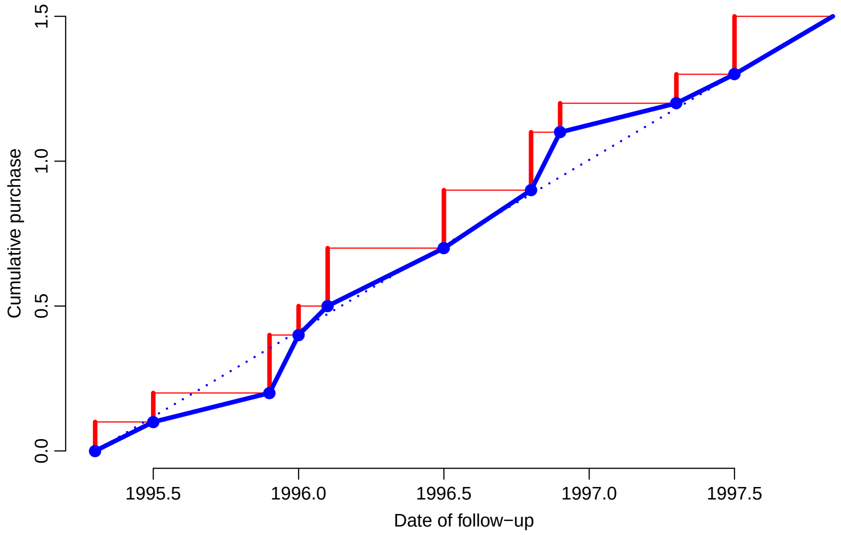

Figure 5.2: *Cumulative purchased dose (red line), and assumed cumulative ingested dose (blue line). The function* `gen.exp` *computes the value of the blue line at prespecified times if* `dpt=NULL`.

```
+                            dof = dof,
+                        cum.amt = cum.amt ) )
+         } ) )
+ }
```

5.3.3 Comparing approaches

We can compare how the two approaches perform by plotting the two results on top of each other for the three patients in the bogus data set:

```
> zzc <- use.amt.dpt( PUR )
> zzx <- use.only.amt( PUR )
> par( mar=c(3,3,1,1), mgp=c(3,1,0)/1.6 )
> plot( NA, xlim=floor(range(zzc$dof))+0:1,
+           ylim=c(0,max(zzc$cum.amt)),
+           xlab="Date", ylab="Cumulative dose", bty="n" )
> for( i in 1:3 )
+    {
+    with( subset(zzc,id==i), lines(dof,cum.amt,col=i+1,lwd=2,lty=1,type="b") )
+    with( subset(zzx,id==i), lines(dof,cum.amt,col=i+1,lwd=2,lty=1,type="b",pch=16) )
+    }
```

5.3.4 Putting it together

Now we have records for the entire follow-up, with an exposure intensity (possibly 0) attached to all intervals. This enables us to compute for example the cumulative dose at the start of each of the intervals, but we are actually not interested in the cumulative dose at the start of a set of analysis intervals.

In practical settings we want to compute the exposures at a set of prespecified times, which typically will be calendar time points. They are supplied in the argument `breaks` to the function `gen.exp`.

The central part of the function is calling the linear interpolation function `approx`. Most of the other paraphernalia is finding the subset of the `breaks` which are relevant for a particular person.

The first part of the function here is is just sorting out which of the two work-horses `use.amt.dpt` and `use.only.amt` to use:

```
> gen.exp <-
+ function( purchase, id="id", dop="dop", amt="amt", dpt="dpt",
+                 fu, doe="doe", dox="dox",
+             breaks,
+            use.dpt = ( dpt %in% names(purchase) ),
+               lags = NULL,
+           push.max = Inf,
+           pred.win = Inf,
+            lag.dec = 1 )
+ {
+ # Make sure that the data fames have the right column names
+ wh <- match( c(id,dop,amt), names(purchase) )
+ if( any( is.na(wh) ) ) stop("Wrong column names for the purchase data frame")
+ names( purchase )[wh] <- c("id","dop","amt")
+ wh <- match( c(id,doe,dox), names(fu) )
+ if( any( is.na(wh) ) ) stop("Wrong column names for the follow-up data frame")
+ names( fu )[wh] <- c("id","doe","dox")
+
+ if( use.dpt )
+   {
+   # This is to allow dpt to be entered as numerical scalar common for all
+   if( is.numeric(dpt) )
+     {
+     if( length(dpt) > 1 ) stop( "If dpt is numeric it must have lenght 1" )
+     purchase$dpt <- dpt
+     }
+   else
+   names( purchase )[match(c(dpt),names(purchase))] <- "dpt"
+   tmp.dfr <- Epi:::use.amt.dpt( purchase,
+                                  lags = lags,
```

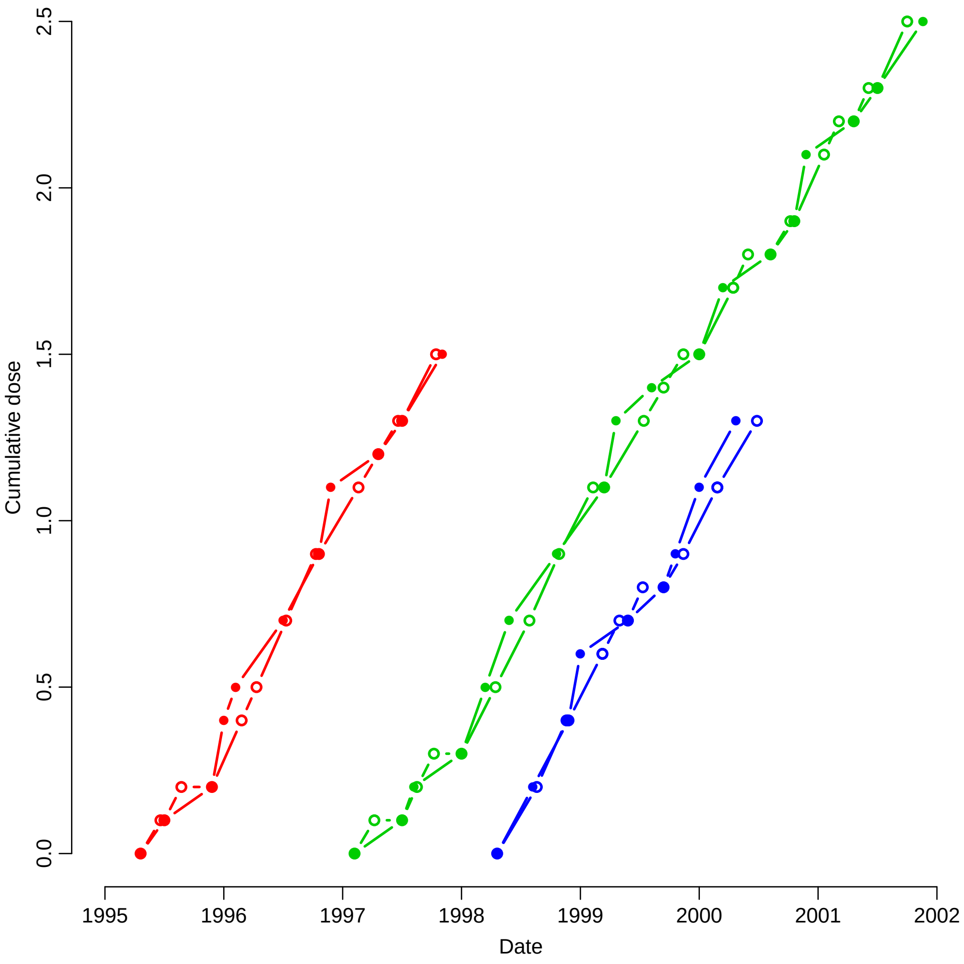

Figure 5.3: *The two approaches to evaluation of cumulative dose: the lines with open symbols use the drug intensity information, and so normally will have the exposure later than the approach (filled symbols) that lets all exposure start at the date of purchase. Normally the two approaches will yield the same eventual cumulative dose. The exception is if the parameter* ***push.max*** *is used, in which case some (part of some) drug purchases will be deemed non-consumed.*

```
+                                  push.max = push.max,
+                                   lag.dec = lag.dec )
+   }
+ else
+   tmp.dfr <- Epi:::use.only.amt( purchase,
+                                       lags = lags,
+                                   pred.win = pred.win,
+                                    lag.dec = lag.dec )
+
+
+ # Merge in the follow-up period for the persons
+ tmp.dfr <- merge( tmp.dfr, fu, all=T )
+
+ # Interpolate to find the cumulative doses at the dates in breaks
+ do.call( "rbind",
+ lapply( split( tmp.dfr, tmp.dfr$id ),
+         function(set)
+         {
+         # All values of these are identical within each set (=person)
+         doe <- set$doe[1]
+         dox <- set$dox[1]
+         # The first and last date of exposure according to the assumption
+         doi <- min(set$dof)
+         doc <- max(set$dof)
+         # Get the breakpoints and the entry end exit dates
+         breaks <- sort( unique( c(breaks,doe,dox) ) )
+         xval   <- breaks[breaks>=doe & breaks<=dox]
+         dfr    <- data.frame( id = set$id[1],
+                              dof = xval )
+         dfr$tfi  <- pmax(0,xval-doi)
+         dfr$tfc  <- pmax(0,xval-doc)
+         dfr$cdos <- approx( set$dof, set$cum.amt, xout=xval, rule=2 )$y
+         for( lg in lags )
+            dfr[,paste( "ldos",
+                        formatC(lg,format="f",digits=lag.dec),
+                        sep="." )] <-
+            approx( set$dof, set$cum.amt, xout=xval-lg, rule=2 )$y
+         dfr
+         } ) )
+ }
```

It is now easy to plot the trajectories of cumulative dose for each person:

```
> resA <- gen.exp( PUR, fu=fu, breaks=seq(1990,2020,0.5), lags=1:2 )
> resD <- gen.exp( PUR, fu=fu, breaks=seq(1990,2020,0.5), lags=1:2, use.dpt=FALSE )
> str(resA)
```

```
'data.frame':        47 obs. of  7 variables:
 $ id      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ dof     : num  1994 1994 1995 1996 1996 ...
 $ tfi     : num  0 0 0 0.2 0.7 ...
 $ tfc     : num  0 0 0 0 0 ...
 $ cdos    : num  0 0 0 0.1 0.28 ...
 $ ldos.1.0: num  0 0 0 0 0 ...
 $ ldos.2.0: num  0 0 0 0 0 ...
```

```
> plot( resA$dof, resA$cdos, type="n", xlab="Date", ylab="Cumulative dose" )
> for( i in 1:3 )
+    matlines( resA[resA$id==i,2],
+               resA[resA$id==i,-c(1:4)], lwd=2, col=i+1, lty=1, pch=16 )
> for( i in 1:3 )
+    matlines( resD[resD$id==i,2],
+               resD[resD$id==i,-c(1:4)], lwd=2, col=i+1, lty=2, pch=16 )
```

5.3.5 Conditioning on the future?

When we make the calculation of the dose-intensity (i.e. the rate of ingestion) for each purchase we are essentially conditioning on the future, because we can only know the rate of consumption (ingestion) of the drug purchased at a given date if we also know the date of the *next* purchase. Hence, the only formally valid cumulative exposure variables computed this way are those with a lag larger than the largest gap between two successive purchases.

As it is seen from the figure 5.4, there is a very strong correlation between the variables with different lags. In particular, the actual "borrowing" of information from the future is quite limited, and in practice, we would mostly use a lag of at least 1 year.

5.4 Wrapping it all up in R

The previous developments indicates that in order to create the relevant variables from exposure (*i.e.* purchase) records, we need the following input from each drug (group):

- Purchase records with:

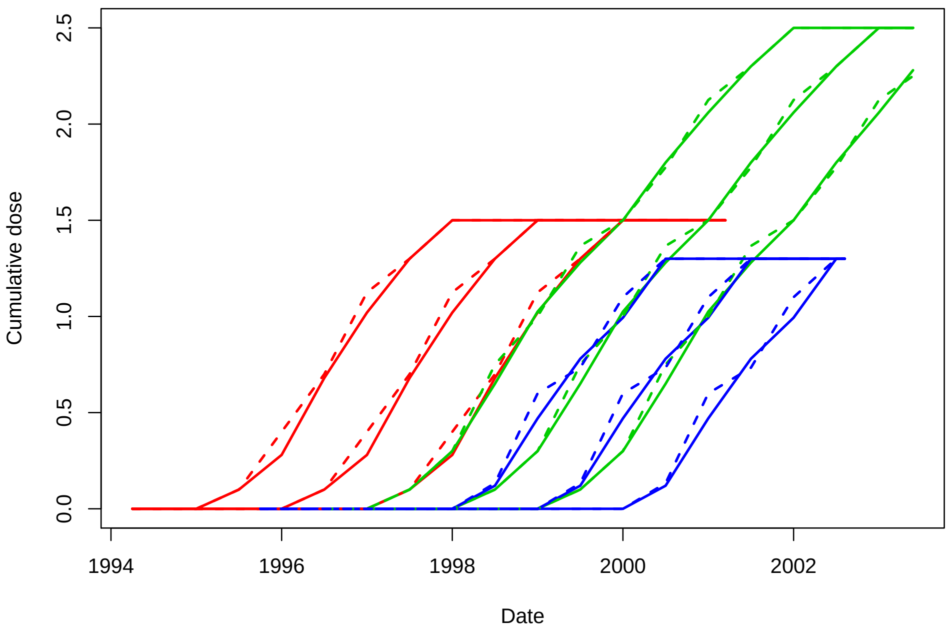

Figure 5.4: *Cumulative dose and lagged cumulative dose for three patients.*

  - Person-id
  - Date of purchase
  - Amount purchased
  - Daily dose for purchase, used to compute the time covered by the purchase (if possible and desirable).

- Entry and exit dates for each person
- A sequence of dates for which we compute the values of the following covariates:
  - Time since first exposure to the drug
  - Time since latest cessation of the drug
  - Cumulative time on the drug (in the absence of dosage information)
  - Cumulative dose of the drug
  - Cumulative dose, lagged
- The lag-times to use for the particular drug in question.

Thus an R-function doing this task should therefore be defined like this:

```
> gen.exp <-
+ function( purchase, id="id", dop="dop", amt="amt", dpt="dpt",
+                 fu, doe="doe", dox="dox",
+             breaks,
+            use.dpt = ( dpt %in% names(purchase) ),
+               lags = NULL,
+           push.max = Inf,
+           pred.win = Inf,
+            lag.dec = 1 ){...}
```

where the arguments are as follows:

`purchase` Data frame of purchases with the following variables (the names of which which can be optionally changed by using the corresponding arguments):

`id` Id of the persons

`dop` Date of purchase

`amt` The dose bought (“amount”)

`dpt` Dose per time. If scalar numeric, it is the `dpt` for all purchases. In units corresponding to `amt/diff(dop)`.

`fu` Data frame of follow-up periods for persons. Multiple records per person are not allowed. The variables are:

`id` Id of the persons

`doe` Date of entry, numeric, same scale as `dop`.

`dox` Date of exit, numeric, same scale as `dop`.

`breaks` A vector of dates at which covariates are computed, same scale as `dop`.

`lags` A (possibly empty) vector of lag-times for computation of lag times for the cumulative dose.

`push.max` Numerical constant. The maximal time that a given purchase can be pushed forward before consumption.

`pred.win` Numerical constant. The length of the time window before the last purchase used to compute the average dose rates which is used as consumption rate fro the last recorded purchase.

`lag.dec` Number of decimals used in annotation of the lagged exposure variables.

The result of the function should be a data frame with columns “`id`”, “`dof`” (date of follow-up; the start of the interval), “`tfi`”, “`tfc`”, “`cdos`”, “`ldos.1.0`”, “`ldos.1.2`”,... where the last two represent the cumulative doses at 1.0, 1.2, ... prior to the follow-up date in `dof`.

If we have date of birth (`dob`) and date of diagnosis of diabetes (`doDM`) available, we can compute the current age as `dof-dob`, and the duration of disease as `dof-doDM`.

5.4.1 The actual function

Based on the code above the function is included in the `Epi`-package, and it looks like this: The functionality of this function is illustrated here, using the same data as we used to develop it.

```
> xpos <- gen.exp( PUR,
+                   fu = fu,
+               breaks = seq(1990,2015,0.2),
+                 lags = 1:4/5 )
> cbind( id=xpos[1:20,1], round( xpos[1:20,-1], 3 ) )
```

```
     id     dof tfi   tfc cdos ldos.0.2 ldos.0.4 ldos.0.6 ldos.0.8
1.1   1 1994.25 0.0 0.000 0.00     0.00     0.00     0.00     0.00
1.2   1 1994.40 0.0 0.000 0.00     0.00     0.00     0.00     0.00
1.3   1 1994.60 0.0 0.000 0.00     0.00     0.00     0.00     0.00
1.4   1 1994.80 0.0 0.000 0.00     0.00     0.00     0.00     0.00
1.5   1 1995.00 0.0 0.000 0.00     0.00     0.00     0.00     0.00
1.6   1 1995.20 0.0 0.000 0.00     0.00     0.00     0.00     0.00
1.7   1 1995.40 0.1 0.000 0.06     0.00     0.00     0.00     0.00
1.8   1 1995.60 0.3 0.000 0.17     0.06     0.00     0.00     0.00
1.9   1 1995.80 0.5 0.000 0.20     0.17     0.06     0.00     0.00
1.10  1 1996.00 0.7 0.000 0.28     0.20     0.17     0.06     0.00
1.11  1 1996.20 0.9 0.000 0.44     0.28     0.20     0.17     0.06
1.12  1 1996.40 1.1 0.000 0.60     0.44     0.28     0.20     0.17
1.13  1 1996.60 1.3 0.000 0.76     0.60     0.44     0.28     0.20
1.14  1 1996.80 1.5 0.000 0.90     0.76     0.60     0.44     0.28
1.15  1 1997.00 1.7 0.000 1.02     0.90     0.76     0.60     0.44
1.16  1 1997.20 1.9 0.000 1.14     1.02     0.90     0.76     0.60
1.17  1 1997.40 2.1 0.000 1.26     1.14     1.02     0.90     0.76
1.18  1 1997.60 2.3 0.000 1.37     1.26     1.14     1.02     0.90
1.19  1 1997.80 2.5 0.014 1.50     1.37     1.26     1.14     1.02
1.20  1 1998.00 2.7 0.214 1.50     1.50     1.37     1.26     1.14
```

We can then plot the values of the covariates for each of the follow-up points (*i.e.* at the start of each of the intervals):

```
> # How many relevant columns ?
> nvar <- ncol(xpos)-3
> clrs <- rainbow(nvar)
> # Show how the variables relate to the follow-up time
> par( mfrow=c(3,1), mar=c(3,3,1,1), mgp=c(3,1,0)/1.6, bty="n" )
> for( i in unique(xpos$id) )
+ matplot( xpos[xpos$id==i,"dof"],
+          xpos[xpos$id==i,-(1:3)],
+          xlim=range(xpos$dof), ylim=range(xpos[,-(1:3)]),
+          type="l", lwd=2, lty=1, col=clrs,
+          ylab="", xlab="Date of follow-up" )
> # Position the variable names
> ytxt <- par("usr")[3:4]
> ytxt <- ytxt[1] + (nvar:1)*diff(ytxt)/(nvar+2)
> xtxt <- rep( sum(par("usr")[1:2]*c(0.98,0.02)), nvar )
> text( xtxt, ytxt, colnames(xpos)[-(1:3)], font=2,
+                   col=clrs, cex=1.5, adj=0 )
```

In inspection of figure 5.5 reveals that the defined set of time-dependent covariates are strongly correlated. In practical modeling it will therefore be difficult to accommodate more that one of these. We shall return to this later.

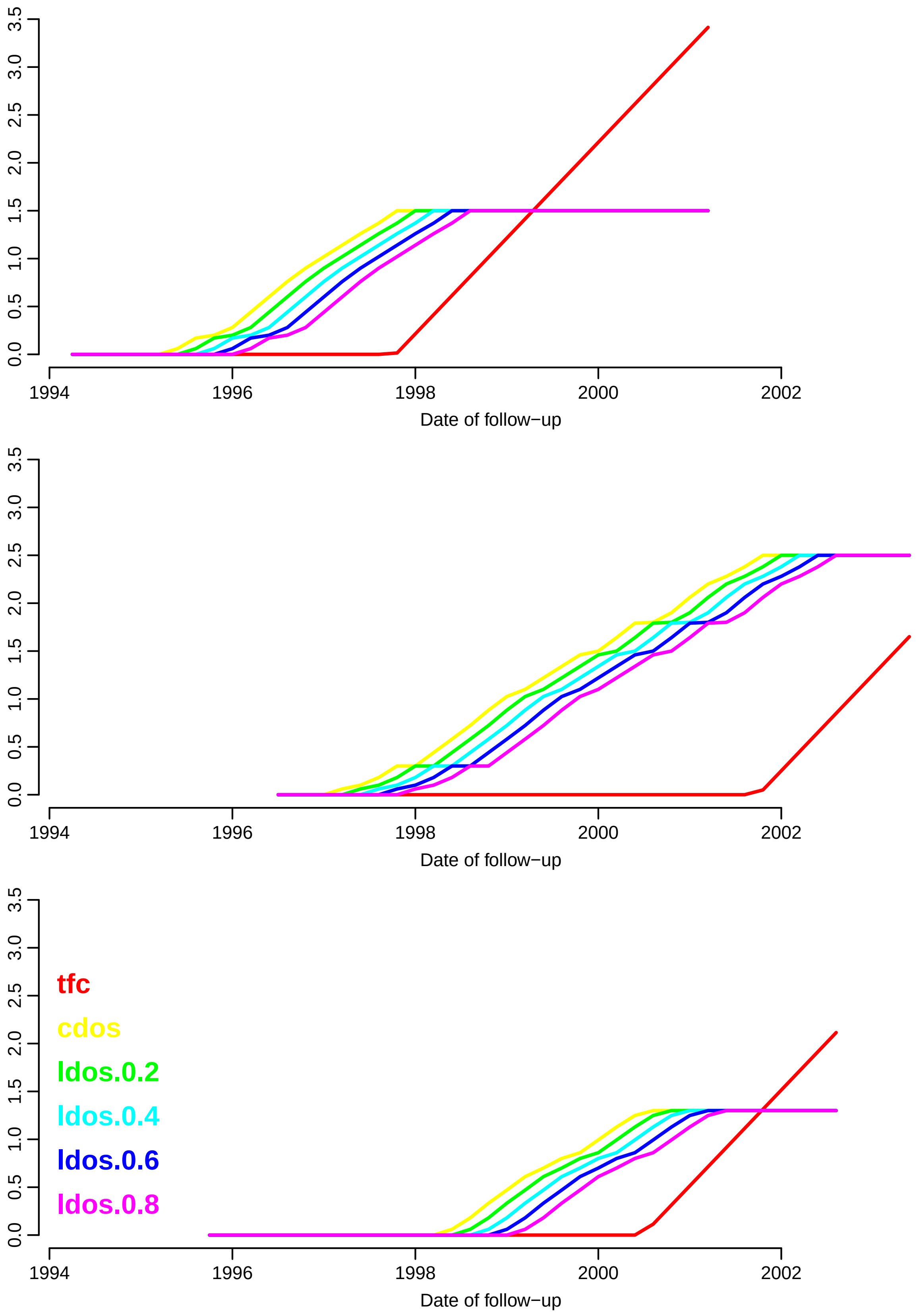

Figure 5.5: *Values of the the defined covariates for the three persons in* `dfr/fu`*, as a function of calendar time.*

5.5 The function documentation for gen.exp

The function is available as part of the Epi package, but not yet (as of Tuesday 3rd January, 2012) on CRAN (**C**omprehensive **R** **A**rchive **N**etwork, http://cran.r-project.org/), but available from R-forge, directly installable by:

```
> install.packages("Epi", repos="http://R-Forge.R-project.org")
```

gen.exp *Generate covariates for drug-exposure follow-up from drug purchase records.*

Description

From records of drug purchase and possibly known treatment intensity, the time since first drug use and cumulative dose at prespecified times is computed. Optionally, lagged exposures are computed too, i.e. cumulative exposure a prespecified time ago.

Usage

```
gen.exp(purchase, id = "id", dop = "dop", amt = "amt", dpt = "dpt",
              fu, doe = "doe", dox = "dox",
          breaks,
         use.dpt = ( dpt %in% names(purchase) ),
            lags = NULL,
        push.max = Inf,
        pred.win = Inf,
         lag.dec = 1 )
```

Arguments

purchase	Data frame with columns id-person id, dop-date of purchase, amnt-amount purchased, and optionally dpt-defined daily dose, that is how much is assumed to be ingested per unit time. The time unit used here is assumed to be the same as that used in dop, so despite the name it is not necessarily measured per day.
id	Name of the id variable in the data frame.
dop	Name of the date of purchase variable in the data frame.
amt	Name of the amount purchased variable in the data frame.
dpt	Name of the dose-per-time variable in the data frame.
fu	Data frame with follow-up period for each person, the person id variable must have the same name as in the purchase data frame.
doe	Name of the date of entry variable.
dox	Name of the date of exit variable.
use.dpt	Logical, should we use information on dose per time.

`breaks`	Numerical vector of time points where the time since exposure and the cumulative dose are computed.
`lags`	Numerical vector of lag-times used in computing lagged cumulative doses.
`push.max`	How much can purchases maximally be pushed forward in time. See details.
`pred.win`	The length of the window used for constructing the average dose per time used to compute the duration of the last purchase
`lag.dec`	How many decimals to use in the construction of names for the lagged exposure variables

Details

Each purchase record is converted into a time-interval of exposure.

If `use.dpt` is `TRUE` then the dose per time informatin is used to compute the exposure interval associated with each purchase. Exposure intervals are stacked, that is each interval is put after any previous. This means that the start of exposure to a given purchase can be pushed into the future. The parameter `push.max` indicates the maximally tolerated push. If this is reached by a person, the assumption is that some of the purchased drug is not counted in the exposure calculations.

The `dpt` can either be a constant, basically translating the purchased amount into exposure time the same way for all persons, or it can be a vector with different treatment intensities for each purchase. In any case the cumulative dose is computed taking this into account.

If `use.dpt` is `FALSE` then the exposure from one purchase is assumed to stretch over the time to the next purchase, so we are effectively assuming different rates of dose per time between any two adjacent purchases. Moreover, with theis approach, periods of non-exposure does not exist.

The intention of this function is to generate covariates for a particular drug for the entire follow-up of each person. The reason that the follow-up prior to drug purchase and post-exposure is included is that the covariates must be defined for these periods too, in order to be useful for analysis of disease outcomes.

Value

A data frame with one record per follow-up interval between `breaks`, with columns:

`id` person id.

`dof` date of follow up, i.e. start of interval. Apart from possibly the first interval for each person, this will assume values in the set of the values in `breaks`.

`Y` the length of interval.

`tfi` time from first initiation of drug.

`tfc` time from latest cessation of drug.

`cdur` cumulative time on the drug.

`cdos` cumulative dose.

`ldos` suffixed with one value per element in `lags`, the latter giving the cumulative doses `lags` before `dof`.