Years of Life Lost to Diabetes

LEAD symposium / EDEG 2017
6th May 2017, Dubrovnik, Croatia
http://bendixcarstensen.com/Epi/Courses/EDEG2017
Version 3

Compiled Friday 28th April, 2017, 11:26
from: /home/bendix/teach/AdvCoh/courses/EDEG.2017/pracs/notes.tex

Bendix Carstensen Steno Diabetes Center Copenhagen, Gentofte, Denmark
& Department of Biostatistics, University of Copenhagen
bcar0029@regionh.dk b@bxc.dk
http://BendixCarstensen.com

R-helpers (SDCC): Pernille Falberg Rønn
Gregers Andersen
Dorte Vistisen
Stine Byberg
Hanan Amadid

Contents

Chapter 0

Introduction

This is a set of notes put together for the LEAD Symposium on May 6th, preceding the EDEG 2017 meeting in Dubrovnik, Croatia.

The first chapter is an explanation of the concept of life lost to a disease, with an emphasis on the necessary probability theory, and a description of the tools for calculation of life lost available in the Epi package for R.

The second chapter is a practical exercise that you are supposed to work through (with appropriate help from the faculty). It falls in two parts; the first is a simple case based on a set of empirical mortality and diabetes incidence rates; the second part of the exercise chapter is an illustration of how to use modeling to get a more detailed and credible picture of the changes in the years of life lost over the last 20 years in Denmark.

The third chapter is merely a reference card of the essential concepts used in description of occurrence rates in cohort studies, notably in population-wide studies.

Symposium timetable

10:00–10:10	Introductions (Dorte Vistisen)
10:10–11:00	Years of life lost (BxC): — prerequisites, assumptions definitions — introduction to practical calculations
11:00–12:00 –	Practicals (BxC + faculty) Based on “real” Danish data you will do calculations of YLL using R and end up with estimates of years of life lost to diabetes in Denmark.
12:00–12:15	Coffee break
12:15–12:45	Review of practicals (BxC)

Installing R and Epi

The symposium contains a practical that requires you to use R, hence you should install R on your computer; get it at https://cran.r-project.org/, you may want to use R-studio which is also free: https://www.rstudio.com/. Finally, once you have started R you should install the Epi package, for example by writing the following at the command line::

```
> install.packages( "Epi" )
> library( Epi)
> sessionInfo()
```

The last command should show you that you have the latest version of R (version 3.4.0) and version 2.12 of the `Epi` package.

If you already have `Epi` installed you may want to update it if you do not have version 2.12, by doing:

```
> update.packages( "Epi" )
> library( Epi)
> sessionInfo()
```

Chapter 1

Survival and years of life lost

This chapter contains a lot of integrals, and integrals are known to be scary or mysterious to many. But you should keep in mind that integrals are just a convenient notation for the area under a curve. The curve being the function inside the integration. This is explained briefly at the start of the exercises, so do not dispair.

1.1 Years of life lost (YLL)

The general concept in calculation of “years lost to. . . ” is the comparison of the expected lifetime between two groups of persons; one without and one with disease (in this example DM). The expected lifetime is the area under the survival curve, so basically the exercise requires that two survival curves that are deemed relevant be available.

The years of life lost is therefore just the area between the survival curves for those “Well”, $S_W(t)$, and for those “Diseased”, $S_D(t)$:

$$\text{YLL} = \int_0^\infty S_W(t) - S_D(t) \, \mathrm{d}t$$

The time t could of course be age, but it could also be “time after age 50” and the survival curves compared would then be survival curves *conditional* on survival till age 50, and the YLL would be the years of life lost for a 50-year old person with diabetes as comapred to a 50-year old person without.

If we are referring to the expected lifetime we will more precisely use the label expected residual life time, ERL.

1.2 Constructing the survival curves

The survival fora person aged 50, say, with diabetes is computed from the mortality rates for persons wth diabetes in ages 50 or more.

YLL can be computed in two different ways, depending on the way the survival curve and hence the expected lifetime of a person *without* diabetes is computed:

- Assume that the “Well” persons are *immune* to disease — using only the non-DM mortality rates throughout for calculation of expected life time.

- Assume that the "Well" persons *can* acquire the disease and thereby see an increased mortality, thus involving all three rates shown in figure 1.1.

The former gives a higher YLL because the comparison is to persons assumed immune to DM — it is assumed that they will never move to the DM state and see a higher mortality. The latter gives a more realistic picture of the comparison of group of persons with and without diabetes at a given age that can be interpreted in the real world.

The differences can be illustrated by figure 1.1; the immune approach corresponds to an assumption of $\lambda(t) = 0$ in the calculation of the survival curve for a person in the "Well" state.

Calculation of the survival of a diseased person already in the "DM" state is unaffected by assumptions about λ.

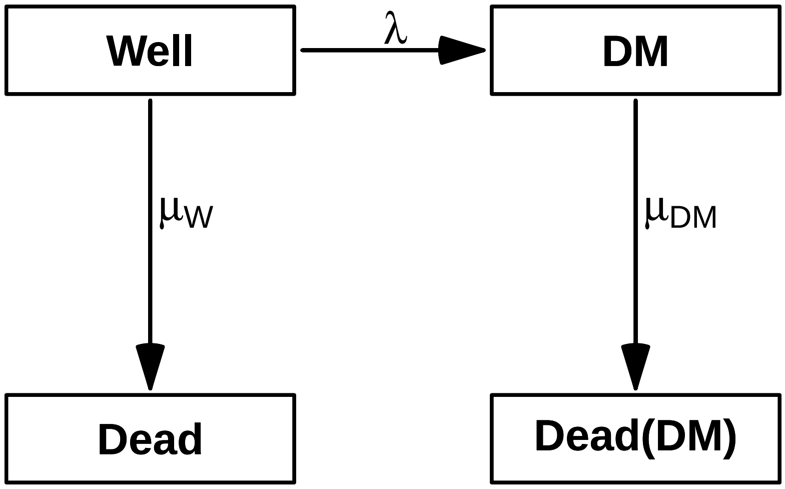

Figure 1.1: *Illness-death model describing diabetes incidence and -mortality in a population.*

1.2.1 Total mortality — a shortcut?

A practical crude shortcut could be to compare the ERL in the diabetic population to the ERL for the *entire* population (that is use the total mortality ignoring diabetes status).

Note however that this approach also counts the mortality of persons that acquired the disease earlier, thus making the comparison population on average more ill than the population we aim at, namely those well at a given time, which only then become more gradually ill.

How large these effects are must be empirically explored, as we shall do later.

1.2.2 Disease duration

In the exposition above there is no explicit provision for the effect of disease duration, but if we were able to devise mortality rates for any combination of age and duration, this could be taken into account.

There are however severe limitations in this as we in principle would want to have duration effects as long as the age-effects — in principle for all combinations (a, d) where $d \leq 100 - A$, where A is the age at which we condition. So even if we were only to compute ERL from age, say, 40 we would still need duration effects up to 60 years (namely to age 100).

The incorporation of duration effects is in principle trivial from a computational point of view, but we would be forced to entertain models predicting duration effects way beyond what is actually observed disease duration in any practical case.

1.2.3 Computing integrals

The practical calculations of survival curves, ERL and YLL involves calculation of (cumulative) integrals of rates and functions of these as we shall see below. This is easy if we have a closed form expression of the function, so its value may be computed at any time point — this will be the case if we model rates by smooth parametric functions.

Computing the (cumulative) integral of a function is done as follows:

- Compute the value of the function (mortality rate for example) at the midpoints of a sequence of narrow equidistant intervals — for example one- or three month intervals of age, say.
- Take the cumulative sum of these values multiplied by the interval length — this will be a very close approximation to the cumulative integral evaluated at the end of each interval.
- If the intervals are really small (like 1/100 year), the distinction between the value at the middle and at the end of each interval becomes irrelevant.

Note that in the above it is assumed that the rates are given in units corresponding to the interval length — or more precisely, as the cumulative rates over the interval.

1.3 Survival functions in the illness-death model

The survival functions for persons in the "Well" state can be computed under two fundamentally different scenarios, depending on whether persons in the "Well" state are assumed to be immune to the disease ($\lambda(a) = 0$) or not.

1.3.1 Immune approach

In this case both survival functions for person in the two states are the usual simple transformation of the cumulative mortality rates:

$$S_W(a) = \exp\left(-\int_0^a \mu_W(u)\,\mathrm{d}u\right), \qquad S_D(a) = \exp\left(-\int_0^a \mu_D(u)\,\mathrm{d}u\right)$$

1.3.1.1 Conditional survival functions

If we want the *conditional* survival functions given survival to age A, say, they are just:

$$S_W(a|A) = S_W(a)/S_W(A), \qquad S_D(a|A) = S_D(a)/S_D(A)$$

1.3.2 Non-immune approach

For a diseased person, the survival function in this states is the same as above, but the survival function for a person without disease (at age 0) is (see figure 1.1):

$$S(a) = \mathrm{P}\{\mathrm{Well}\}(a) + \mathrm{P}\{\mathrm{DM}\}(a)$$

In the appendix of the paper [2] is an indication of how to compute the probability of being in any of the four states shown in figure 1.1, which I shall repeat here:

In terms of the rates, the probability of being in the "Well" box is simply the probability of escaping both death (at a rate of $\mu_W(a)$) and diabetes (at a rate of $\lambda(a)$):

$$\mathrm{P}\{\mathrm{Well}\}(a) = \exp\left(-\int_0^a \mu_W(u) + \lambda(u)\right)\,\mathrm{d}u$$

The probability of being alive with diabetes at age a, is computed given that diabetes occurred at age s ($s < a$) and then integrated over s from 0 to a:

$$\begin{aligned}\mathrm{P}\{\mathrm{DM}\}(a) &= \int_0^a \mathrm{P}\{\text{survive to } s,\ \text{DM diagnosed at } s\} \\ &\qquad \times \mathrm{P}\{\text{survive with DM from } s \text{ to } a\}\,\mathrm{d}s \\ &= \int_0^a \lambda(s)\exp\left(-\int_0^s \mu_W(u) + \lambda(u)\,\mathrm{d}u\right) \\ &\qquad \times \exp\left(-\int_s^a \mu_D(u)\,\mathrm{d}u\right)\,\mathrm{d}s\end{aligned}$$

Sometimes we will use a version where the mortality among diabetes patients depend both on age a and duration of diabetes, d, $\mu_D(a,d)$, in which case we get:

$$\begin{aligned}\mathrm{P}\{\mathrm{DM}\}(a) = \int_0^a \lambda(s)\exp\left(-\int_0^s \mu_W(u) + \lambda(u)\,\mathrm{d}u\right) \\ \times \exp\left(-\int_s^a \mu_D(u, u-s)\,\mathrm{d}u\right)\,\mathrm{d}s\end{aligned}$$

because the integration variable u is the age-scale and the second integral refers to mortality among persons diagnosed at age s, that is, with duration $u - s$ at age u.

The option of using duration-dependent mortality rates among diseased individuals is not implemented yet.

1.3.2.1 Conditional survival functions

Unlike the immune approach, the conditional survival function in the more realistic case is not just a ratio of the unconditional survival function to its value at the conditioning age, A, say. This would amount to conditioning on being merely *alive* at age A, but what we want is to condition on being in the "Well" state at age A.

The formulae for the conditional probabilities of being either in "Well" or "DM", given being in "Well" at age A are basically replicates of the unconditional, albeit with changes in integration limits:

$$
\begin{aligned}
\mathrm{P}\{\text{Well}|\text{Well at } A\}\,(a) &= \exp\left(-\int_A^a \mu_W(u)+\lambda(u)\right)\,\mathrm{d}u \\
\mathrm{P}\{\text{DM}|\text{Well at } A\}\,(a) &= \int_A^a \lambda(s)\exp\left(-\int_A^s \mu_W(u)+\lambda(u)\,\mathrm{d}u\right) \\
&\qquad \times \exp\left(-\int_s^a \mu_D(u,u-s)\,\mathrm{d}u\right)\,\mathrm{d}s
\end{aligned}
$$

The calculation of these conditional survival functions is implemented but not allowing for duration-dependence. Thus it is only implemented assuming $\mu_D(a,d)=\mu_D(a)$.

1.4 Practical implementation

There are functions that wraps these formulae up for practical use, available in the Epi package — here is a printout of the documentation:

`erl` *Compute survival functions from rates and expected residual lifetime in an illness-death model as well as years of life lost to disease.*

Description

These functions compute survival functions from a set of mortality and disease incidence rates in an illness-death model. Expected residual life time can be computed under various scenarios by the `erl` function, and areas between survival functions can be computed under various scenarios by the `yll` function. Rates are assumed supplied for equidistant intervals of length `int`.

Usage

```
surv1( int, mu ,                 age.in = 0, A = NULL )
 erl1( int, mu ,                 age.in = 0 )
surv2( int, muW, muD, lam,       age.in = 0, A = NULL )
  erl( int, muW, muD, lam=NULL, age.in = 0, A = NULL,
       immune = is.null(lam), yll=TRUE, note=TRUE )
  yll( int, muW, muD, lam=NULL, age.in = 0, A = NULL,
       immune = is.null(lam), note=TRUE )
```

Arguments

`int` Scalar. Length of intervals that rates refer to.

`mu` Numeric vector of mortality rates at midpoints of intervals of length `int`

`muW` Numeric vector of mortality rates among persons in the ”Well” state at midpoints of intervals of length `int`. Left endpoint of first interval is `age.in`.

`muD` Numeric vector of mortality rates among persons in the ”Diseased” state at midpoints of intervals of length `int`. Left endpoint of first interval is `age.in`.

`lam` Numeric vector of disease incidence rates among persons in the ”Well” state at midpoints of intervals of length `int`. Left endpoint of first interval is `age.in`.

`age.in` Scalar indicating the age at the left endpoint of the first interval.

`A`	Numeric vector of conditioning ages for calculation of survival functions.
`immune`	Logical. Should the years of life lost to the disease be computed using assumptions that non-diseased individuals are immune to the disease (`lam`=0) and that their mortality is yet still `muW`.
`note`	Logical. Should a warning of silly assumptions be printed?
`yll`	Logical. Should years of life lost be included in the result?

Details

The mortality rates given are supposed to refer to the ages `age.in+(i-1/2)*int`, `i=1,2,3,...`.

The units in which `int` is given must correspond to the units in which the rates `mu`, `muW`, `muD` and `lam` are given. Thus if `int` is given in years, the rates must be given in the unit of events per year.

The ages in which the survival curves are computed are from `age.in` and then at the end of `length(muW)` (`length(mu)`) intervals each of length `int`.

The `age.in` argument is merely a device to account for rates only available from a given age. It has two effects, one is that labeling of the interval endpoint is offset by this quantity, thus starting at `age.in`, and the other that the conditioning ages given in the argument `A` will refer to the ages defined by this.

The `immune` argument is `FALSE` whenever the disease incidence rates are supplied. If set to `TRUE`, the years of life lost is computed under the assumption that individuals without the disease at a given age are immune to the disease in the sense that the disease incidence rate is 0, so transitions to the diseased state (with presumably higher mortality rates) are assumed not to occur. This is a slightly peculiar assumption (but presumably the most used in the epidemiological literature) and the resulting object is therefore given an attribute, `NOTE`, that point this out. The default of the `surv2` function is to take the possibility of disease into account in order to potentially rectify this.

Value

`surv1` and `surv2` return a matrix whose first column is the ages at the ends of the intervals, thus with `length(mu)+1` rows. The following columns are the survival functions (since `age.in`), and conditional on survival till ages as indicated in `A`, thus a matrix with `length(A)+2` columns. Columns are labeled with the actual conditioning ages; if `A` contains values that are not among the endpoints of the intervals used, the nearest smaller interval border is used as conditioning age, and columns are named accordingly.

`surv1` returns the survival function for a simple model with one type of death, occurring at intensity `mu`.

`surv2` returns the survival function for a person in the ”Well” state of an illness-death model, taking into account that the person may move to the ”Diseased” state, thus requiring all three transition rates to be specified. The conditional survival functions are conditional on being in the ”Well” state at ages given in `A`.

`erl1` returns a three column matrix with columns `age`, `surv` (survival function) and `erl` (expected residual life time) with `length(mu)+1` rows.

`erl` returns a two column matrix, columns labeled ”Well” and ”Dis”, and with row-labels `A`. The entries are the expected residual life times given survival to `A`. If `yll=TRUE` the difference between the columns is added as a third column, labeled ”YLL”.

Author(s)

Bendix Carstensen, `<b@bxc.dk>`

See Also

`ci.cum`

Examples

```
library( Epi )
data( DMlate )
# Naive Lexis object
Lx <- Lexis( entry = list( age = dodm-dobth ),
              exit = list( age = dox -dobth ),
       exit.status = factor( !is.na(dodth), labels=c("DM","Dead") ),
              data = DMlate )
# Cut follow-up at insulin inception
Lc <- cutLexis( Lx, cut = Lx$doins-Lx$dob,
              new.state = "DM/ins",
       precursor.states = "DM" )
summary( Lc )
# Split in small age intervals
Sc <- splitLexis( Lc, breaks=seq(0,120,2) )
summary( Sc )

# Overview of object
boxes( Sc, boxpos=TRUE, show.BE=TRUE, scale.R=100 )

# Knots for splines
a.kn <- 2:9*10

# Mortality among DM
mW <- glm( lex.Xst=="Dead" ~ Ns( age, knots=a.kn ),
           offset = log(lex.dur),
           family = poisson,
             data = subset(Sc,lex.Cst=="DM") )

# Mortality among insulin treated
mI <- update( mW, data = subset(Sc,lex.Cst=="DM/ins") )

# Total motality
mT <- update( mW, data = Sc )

# Incidence of insulin inception
lI <- update( mW, lex.Xst=="DM/ins" ~ . )

# From these we can now derive the fitted rates in intervals of 1 year's
# length. In real applications you would use much smaller interval like
# 1 month:
# int <- 1/12
int <- 1

# Prediction frame to return rates in units of cases per 1 year
# - we start at age 40 since rates of insulin inception are largely
# indeterminate before age 40
nd <- data.frame( age = seq( 40+int, 110, int ) - int/2,
              lex.dur = 1 )
```

```
muW <- predict( mW, newdata = nd, type = "response" )
muD <- predict( mI, newdata = nd, type = "response" )
lam <- predict( lI, newdata = nd, type = "response" )

# Compute the survival function, and the conditional from ages 50 resp. 70
s1 <- surv1( int, muD, age.in=40, A=c(50,70) )
round( s1, 3 )

s2 <- surv2( int, muW, muD, lam, age.in=40, A=c(50,70) )
round( s2, 3 )

# How much is YLL overrated by ignoring insulin incidence?
round( YLL <- cbind(
yll( int, muW, muD, lam, A = 41:90, age.in = 40 ),
yll( int, muW, muD, lam, A = 41:90, age.in = 40, immune=TRUE ) ), 2 )[seq(1,51,10),]

par( mar=c(3,3,1,1), mgp=c(3,1,0)/1.6, bty="n", las=1 )
matplot( 40:90, YLL,
         type="l", lty=1, lwd=3,
         ylim=c(0,10), yaxs="i", xlab="Age" )
```

Chapter 2

Exercises

2.1 Introduction

This is an exercise in R, to show you how to use mortality rates to construct survival curves and how to compute the area between them — the years of life lost to a disease.

The rates we use as basis for the following calculations are derived from a reconstructed version of the NDR, covering the period 1996-01-01 through 2015-12-31.

2.1.1 Integrals in practice

Computing the integrals that we see in the formulae is the same as computing the area under the curves.

Computing the area under the curve ($\int f(x)\,\mathrm{d}x$) is in practice done by subdividing the x-axis in small intervals and calculating the value of the function f at the midpoint of each interval. This is then multiplied by the width of each interval, and the contributions added. Done.

It works out a little simpler if all intervals have the same width; then you can just add the function values at the midpoints and then multiply the sum with the (common) width of the intervals.

Either way, it is equivalent to the use of the so-called "trapezoidal rule".

All you need is therefore a dataset with two variables, `width` and `fval`, each observation representing one interval. The area under the curve (the integral) is then

```
> sum( width * fval )
```

This is what is exploited in the functions `erl` and `yll`, but the important feature to recognize is that you need values of mortality/incidence rates at equidistant points in time (at the midpoints of the intervals).

2.2 Mortality and incidence rates

First load the `Epi` package which has the relevant dataset(s) and epidemiological functions:

```
> library( Epi )
> print( sessionInfo(), l=F )
```

```
R version 3.4.0 (2017-04-21)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 14.04.5 LTS

Matrix products: default
BLAS: /usr/lib/libblas/libblas.so.3.0
LAPACK: /usr/lib/lapack/liblapack.so.3.0

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

other attached packages:
[1] Epi_2.12

loaded via a namespace (and not attached):
 [1] cmprsk_2.2-7      MASS_7.3-45       compiler_3.4.0    Matrix_1.2-6
 [5] plyr_1.8.4        parallel_3.4.0    survival_2.41-3   etm_0.6-2
 [9] Rcpp_0.12.5       splines_3.4.0     grid_3.4.0        numDeriv_2014.2-1
[13] lattice_0.20-33
```

1. We load in the dataset of DM and population mortality and incidence in Denmark, `DMepi`:

```
> data( DMepi )
> str( DMepi )
```

```
'data.frame':        4000 obs. of  8 variables:
 $ sex : Factor w/ 2 levels "M","F": 1 2 1 2 1 2 1 2 1 2 ...
 $ A   : num  0 0 1 1 2 2 3 3 4 4 ...
 $ P   : num  1996 1996 1996 1996 1996 ...
 $ X   : num  1 9 4 7 7 2 6 5 9 4 ...
 $ D.nD: num  28 19 23 19 7 8 8 8 6 7 ...
 $ Y.nD: num  35454 33095 36451 34790 35329 ...
 $ D.DM: num  0 0 0 0 0 0 0 0 0 0 ...
 $ Y.DM: num  0.476 3.877 4.92 7.248 12.474 ...
```

```
> summary( DMepi )
```

```
 sex            A               P              X                 D.nD
 M:2000   Min.   : 0.00   Min.   :1996   Min.   :  0.00   Min.   :   0.0
 F:2000   1st Qu.:24.75   1st Qu.:2001   1st Qu.: 12.00   1st Qu.:  14.0
          Median :49.50   Median :2006   Median : 43.00   Median : 102.5
          Mean   :49.50   Mean   :2006   Mean   : 82.39   Mean   : 238.0
          3rd Qu.:74.25   3rd Qu.:2010   3rd Qu.:138.00   3rd Qu.: 408.0
          Max.   :99.00   Max.   :2015   Max.   :542.00   Max.   :1164.0
      Y.nD               D.DM             Y.DM
 Min.   :   48.78   Min.   :  0.00   Min.   :   0.0
 1st Qu.:17053.30   1st Qu.:  0.00   1st Qu.: 134.5
 Median :32144.99   Median : 10.00   Median : 486.8
 Mean   :26451.65   Mean   : 37.84   Mean   : 839.5
 3rd Qu.:35851.41   3rd Qu.: 67.00   3rd Qu.:1278.4
 Max.   :45254.35   Max.   :215.00   Max.   :5300.9
```

A detailed description of the data can be obtained from:

```
> ?DMepi
```

2. The dataset `DMepi` contains no. of incident DM cases, deaths and person-years for persons with and without diabetes:

   For each combination of sex, age (`A` — 100 levels, 0–99) and period (`P` — 20 levels, 1996–2015) in 1 year groups, we have the person-years in the "Well" (`Y.nD`) and the "DM" (`Y.DM`) states, as well as the number of deaths from these (`D.nD`, `D.DM`) and the number of incident diabetes cases from the "Well" state (`X`), the top of the dataset is:

```
> head( DMepi )

  sex A    P X D.nD     Y.nD D.DM       Y.DM
1   M 0 1996 1   28 35453.65    0  0.4757016
2   F 0 1996 9   19 33094.86    0  3.8767967
3   M 1 1996 4   23 36450.73    0  4.9199179
4   F 1 1996 7   19 34789.99    0  7.2484600
5   M 2 1996 7    7 35328.92    0 12.4743326
6   F 2 1996 2    8 33673.43    0  8.0951403
```

3. In order to compute the years of life lost to diabetes we need the survival functions for persons with and without diabetes. These are derived from age-specific incidence and mortality rates. So we first have a look at the mortality and the incidence rates.

   For the sake of simplicity we first restrict to a singe sex and a particular year, in this illustration we use women and the year 2015, but choose you own.

   To do the calculations we must sort the dataset by the age, `A`. The function `order` tells you in which order you should take the elements of a vector to have it sorted, so `order(A)` is the order in which we want the rows:

```
> w15 <- subset( DMepi, sex=="F" & P==2015 )
> w15 <- w15[order(w15$A),]
> head( w15 )

     sex A    P  X D.nD     Y.nD D.DM      Y.DM
3802   F 0 2015  0    8 27692.48    0  0.000000
3804   F 1 2015  4    2 27558.64    0  3.532512
3806   F 2 2015 10    4 28204.69    0  9.576318
3808   F 3 2015  7    1 28916.24    0 14.725530
3810   F 4 2015  4    3 30704.35    0 13.488022
3812   F 5 2015  7    3 31504.41    0 22.655031
```

4. We can then compute the relevant incidence and mortality rates for women in 2015, including the total population mortality rate. The `transform` function does the job of adding variables to a dataset:

```
> w15 <- transform( w15, mW =        D.nD / Y.nD,
+                        iW =           X / Y.nD,
+                        mD = pmax(0,D.DM / Y.DM,na.rm=TRUE),
+                        mT =  (D.nD+D.DM)/(Y.nD+Y.DM) )
```

   The reason for the `pmax()` construction is that some units have `Y.DM` equal to 0, and hence generate `NA`s for the rates. But we want a 0 and not an NA for those.

```
> str( w15 )
'data.frame':        100 obs. of  12 variables:
 $ sex : Factor w/ 2 levels "M","F": 2 2 2 2 2 2 2 2 2 2 ...
 $ A   : num  0 1 2 3 4 5 6 7 8 9 ...
 $ P   : num  2015 2015 2015 2015 2015 ...
 $ X   : num  0 4 10 7 4 7 10 8 7 17 ...
 $ D.nD: num  8 2 4 1 3 3 2 1 4 1 ...
 $ Y.nD: num  27692 27559 28205 28916 30704 ...
 $ D.DM: num  0 0 0 0 0 0 0 0 0 0 ...
 $ Y.DM: num  0 3.53 9.58 14.73 13.49 ...
 $ mW  : num  2.89e-04 7.26e-05 1.42e-04 3.46e-05 9.77e-05 ...
 $ iW  : num  0 0.000145 0.000355 0.000242 0.00013 ...
 $ mD  : num  0 0 0 0 0 0 0 0 0 0 ...
 $ mT  : num  2.89e-04 7.26e-05 1.42e-04 3.46e-05 9.77e-05 ...
```

```
> summary( w15 )
 sex           A               P              X               D.nD            Y.nD
 M:  0   Min.   : 0.00   Min.   :2015   Min.   :  0.00   Min.   :  1.0   Min.   :  563
 F:100   1st Qu.:24.75   1st Qu.:2015   1st Qu.: 16.00   1st Qu.:  6.0   1st Qu.:20889
         Median :49.50   Median :2015   Median : 35.50   Median : 72.0   Median :32238
         Mean   :49.50   Mean   :2015   Mean   : 76.81   Mean   :212.8   Mean   :27406
         3rd Qu.:74.25   3rd Qu.:2015   3rd Qu.:152.50   3rd Qu.:387.0   3rd Qu.:34814
         Max.   :99.00   Max.   :2015   Max.   :269.00   Max.   :823.0   Max.   :41804
      D.DM             Y.DM              mW                 iW
 Min.   :  0.00   Min.   :   0.0   Min.   :0.000029   Min.   :0.0000000
 1st Qu.:  0.00   1st Qu.: 246.2   1st Qu.:0.000179   1st Qu.:0.0005274
 Median : 12.50   Median : 749.3   Median :0.001809   Median :0.0029587
 Mean   : 42.71   Mean   :1157.1   Mean   :0.035719   Mean   :0.0032738
 3rd Qu.: 78.50   3rd Qu.:2012.0   3rd Qu.:0.017944   3rd Qu.:0.0056458
 Max.   :171.00   Max.   :3560.4   Max.   :0.374763   Max.   :0.0081590
       mD                 mT
 Min.   :0.000000   Min.   :0.0000297
 1st Qu.:0.000000   1st Qu.:0.0001779
 Median :0.007897   Median :0.0020369
 Mean   :0.053352   Mean   :0.0373181
 3rd Qu.:0.036594   3rd Qu.:0.0202001
 Max.   :0.539797   Max.   :0.3854294
```

So we now have 4 new variables, representing incidence rates of DM, mortality rates for person with and without DM and

5. We can plot the four different rates on a log-scale to get an overview:

```
> with( w15, matplot( A, cbind( mW, mD, mT, iW)*1000,
+                      log="y", lwd=3, type="l", lty=1,
+                      col=c("red","blue","limegreen","black") ) )
> text( rep(5,4), 500*0.6^c(3,1,2,4), c("mort Well","mort DM","mort Total","DM inc"),
+                      col=c("red","blue","limegreen","black"), adj=0 )
```

The mortality rates among non-DM persons are slightly smaller than the total population mortality. Another feature is the distinctly wiggly feature of the curves. We shall return to this.

6. We can compute the corresponding survival functions using the `surv1` function, that takes a vector of mortality rates as input, assuming that they refer to midpoints of intervals of the same length (argument `int`), first interval starting from 0.

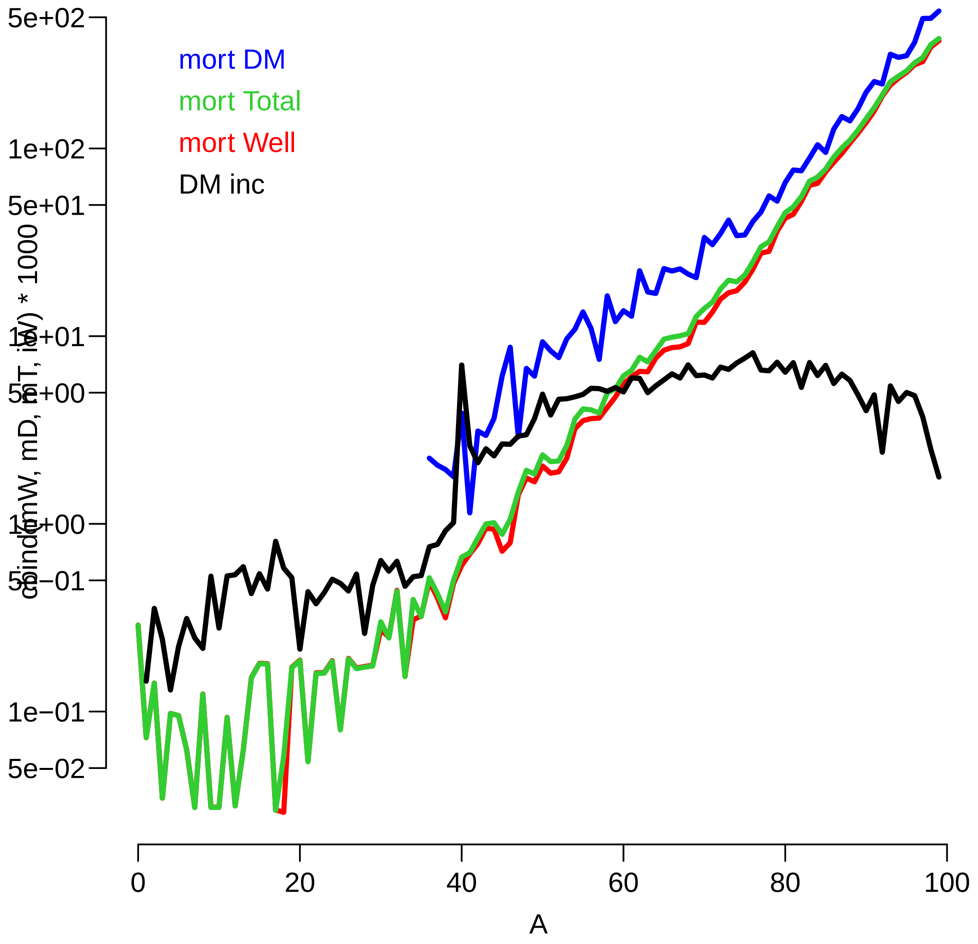

Figure 2.1: *Empirical mortality and incidence rates for women in Denmark 2015.*

```
> with( w15, matplot( surv1( 1, mW )[,1],
+              cbind( surv1( 1, mW )[,2],
+                     surv1( 1, mD )[,2],
+                     surv1( 1, mT )[,2] ),
+                     lwd=3, type="l", lty=1, yaxs="i", ylim=0:1,
+                     col=c("red","blue","limegreen") ) )
```

Among these survival functions, only the green really has a proper interpretation as the survival probability of a person from the general population.

The red curve is the survival of a person without diabetes under the assumptions of 1) diabetes will never occur 2) the mortality rate is the same as among those who can get diabetes during the time before contracting the disease.

Finally, the blue curve is the expected survival of a person with diabetes at birth, assuming that mortality rates do not depend on age at onset or duration of diabetes. The rates used for the calculation are mortality rates for persons with diabetes, regardless of the age in which the person acquired diabetes, so this assumption is highly dubious. If we instead compute *conditional* survival functions, given that the person is, say, 50 years old, we are more likely to get an interpretable survival function.

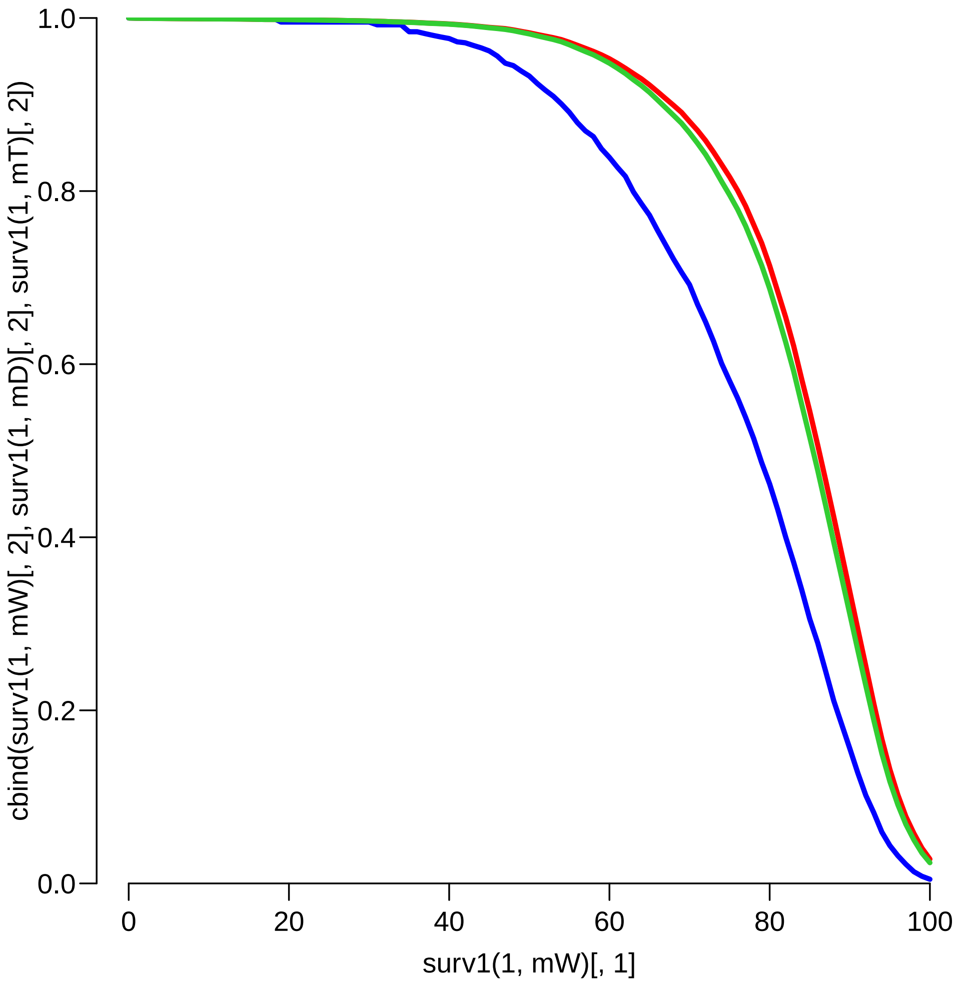

Figure 2.2: *Survival curves for persons with and without DM and for the total population.*

7. We can repeat the exercise for conditional survival given being alive at 50 years:

```
> with( w15, surv1( 1, mW, A=50 ) )[17:23,]

   age        A0 A50
17  16 0.9984714   1
18  17 0.9982920   1
19  18 0.9982621   1
20  19 0.9982331   1
21  20 0.9980611   1
22  21 0.9978735   1
23  22 0.9978192   1

> with( w15, matplot( surv1( 1, mW, A=50 )[,1],
+               cbind( surv1( 1, mW, A=50 )[,3],
+                      surv1( 1, mD, A=50 )[,3],
+                      surv1( 1, mT, A=50 )[,3] ),
+                      lwd=3, type="l", lty=1, yaxs="i", ylim=0:1,
+                      xlab="Age", ylab="Conditional survival given age 50",
+                      col=c("red","blue","limegreen"), xlim=c(50,100) ) )
```

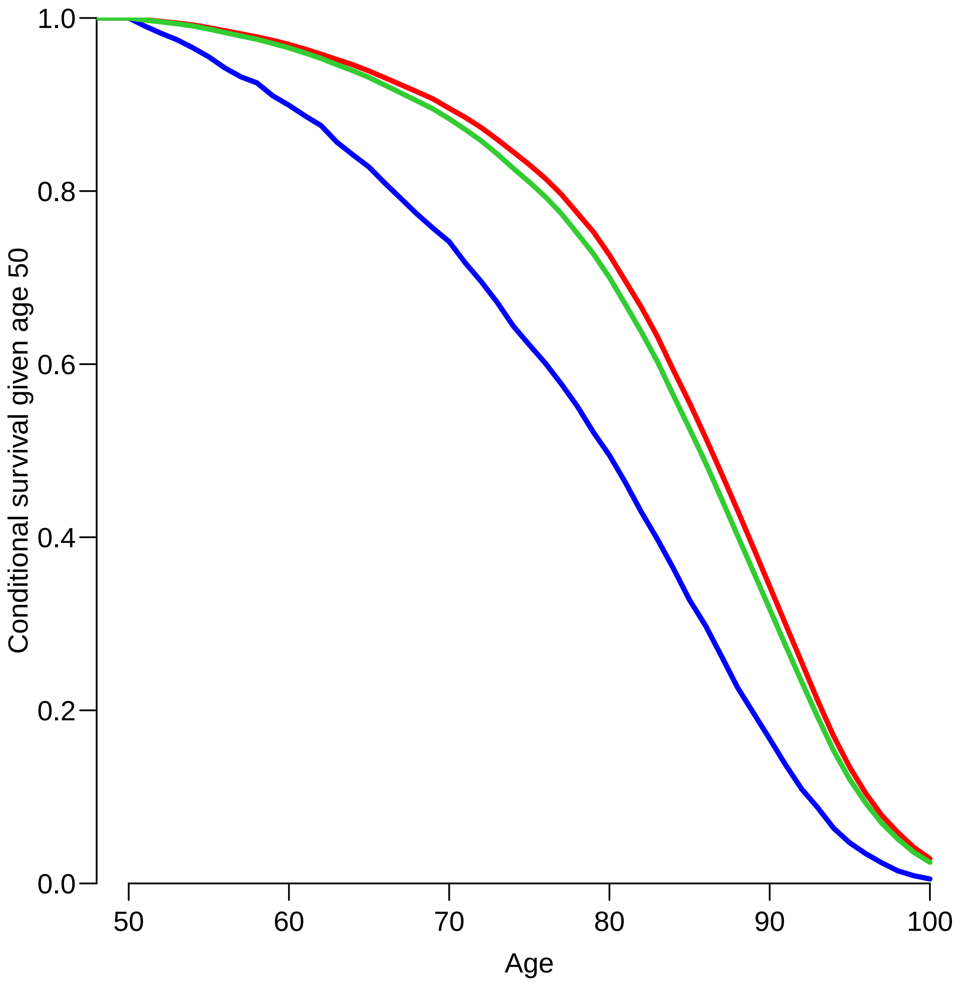

Figure 2.3: *Conditional survival curves. The years of life lost to DM (at age 50) is the area between the blue survival curve for persons with DM and a curve for persons without DM (red, immunity!) or the total population mortality (green).*

8. Finally we compute the survival function for a well person, taking into account the possibility of getting DM on the way and thus have a higher mortality. This rather hairy computation is implemented in the function `surv2` which of course must be supplied with mortality rates for persons with and without diabetes as well as the incidence rates of diabetes.

   We just add this to the previous graph:

```
> with( w15, matplot( surv1( 1, mW, A=50 )[,1],
+               cbind( surv1( 1, mW, A=50 )[,3],
+                      surv1( 1, mD, A=50 )[,3],
+                      surv1( 1, mT, A=50 )[,3],
+                      surv2( 1, mW, mD, iW, A=50 )[,3] ),
+                      lwd=3, type="l", lty=c(1,1,1,2), yaxs="i", ylim=0:1,
+                      xlab="Age", ylab="Conditional survival given age 50",
+                      col=c("red","blue","limegreen","magenta"), xlim=c(50,100) ) )
> text( 95, seq(0.9,0.7,,4), c("","","",""),
+       col=c("red","blue","limegreen","magenta") )
```

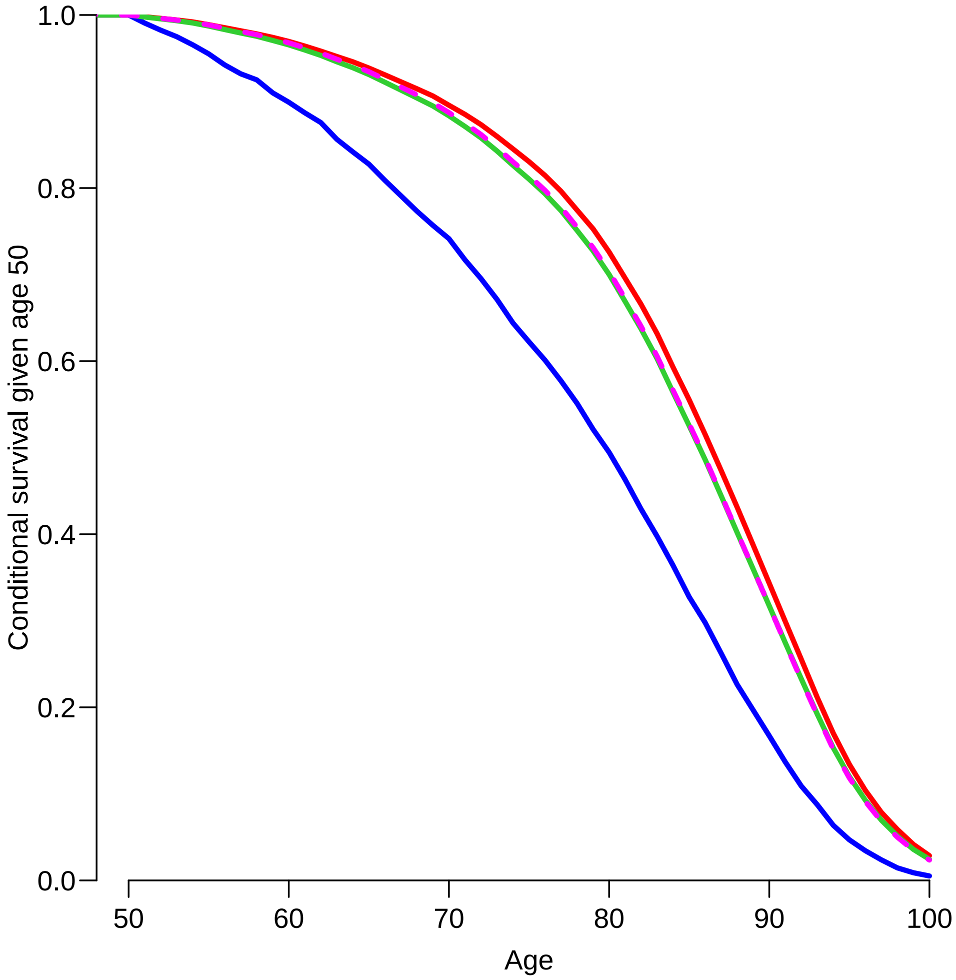

Figure 2.4: *Survival curves conditional on attaining age 50; the added (broken) line is the correctly computed survival function.*

From figure 2.4 we see that the corrected survival is practically indistinguishable from the total population mortality.

2.3 Years of life lost

9. The function `yll` computes the years of life lost under the different scenarios. Since the calculated rates are in events per 1 person-year we can compute the years of life lost, by putting the first argument of `yll` to 1. We also set the `A` argument which indicates the ages at which we want the expected residual life time computed; by default the point 0 is always included — actually it is the value of `age.in` that is used, but the default of this is 0.

```
> with( w15, yll( int=1, muW=mW, muD=mD, lam=iW, A=c(40,50,60,70,80) ) )

       A0       A40       A50       A60       A70       A80 
43.202977  6.787443  5.956740  4.564222  3.168186  1.680120 
```

We can now see how much it matters if we assume immunity and if we replace the mortality among the non-DM with the total mortality, So we omit the `lam` (diabetes incidence) argument, and use either the mortality among persons without DM (immunity assumption) or the total mortality in the population:

```
> with( w15, yll( int=1, muW=mW, muD=mD, lam=iW, A=c(40,50,60,70,80) ) )

       A0       A40       A50       A60       A70       A80 
43.202977  6.787443  5.956740  4.564222  3.168186  1.680120 

> with( w15, yll( int=1, muW=mW, muD=mD, A=c(40,50,60,70,80), n=F ) )

       A0       A40       A50       A60       A70       A80 
44.155298  7.610837  6.584063  4.954874  3.358854  1.739498 

> with( w15, yll( int=1, muW=mT, muD=mD, A=c(40,50,60,70,80), n=F ) )

       A0       A40       A50       A60       A70       A80 
43.399315  6.859584  5.865477  4.333904  2.888800  1.488385 
```

We see that there is a substantial difference between the approaches — the corrected approach taking incidence into account gives results between the two other approaches — the immunity assumption *over*estimates YLL and the uses of the total mortality tends ti *under*estimate the YLL. At lest for this example.

10. To plot how years of life lost looks as a function of age, we use some more tightly spaced ages for conditioning, and we put the results in a vector:

```
> yllf2015 <- with( w15, yll( int=1, muW=mW, muD=mD, lam=iW, A=c(40:90) ) )
> plot( 40:90, yllf2015[-1], type="l", lwd=3 )
```

Now try to see graphically what happens if we ignore the DM incidence and just compare to “immune” persons:

```
> yllf2015x <- with( w15, yll( int=1, muW=mW, muD=mD, A=c(40:90) ) )

NOTE: Calculations assume that Well persons cannot get Ill (quite silly!).

> lines( 40:90, yllf2015x[-1], type="l", lwd=3, lty="22" )
```

Finally compare to the *total* mortality:

```
> yllf2015t <- with( w15, yll( int=1, muW=mT, muD=mD, A=c(40:90), note=F ) )
> lines( 40:90, yllf2015t[-1], type="l", lwd=3, lty="44" )
```

We can redo it using a better scaling of the y-axis (using `ylim=`): `yaxs="i"` makes sure that the lover end of the y-axis is at the levels of the x-axis, so that visual distance from the x-axis actually is the YLL:

```
>  plot( 40:90, yllf2015 [-1], type="l", lwd=3, ylim=c(0,8), yaxs="i" )
> lines( 40:90, yllf2015x[-1], type="l", lwd=3, lty="12" )
> lines( 40:90, yllf2015t[-1], type="l", lwd=3, lty="53" )
```

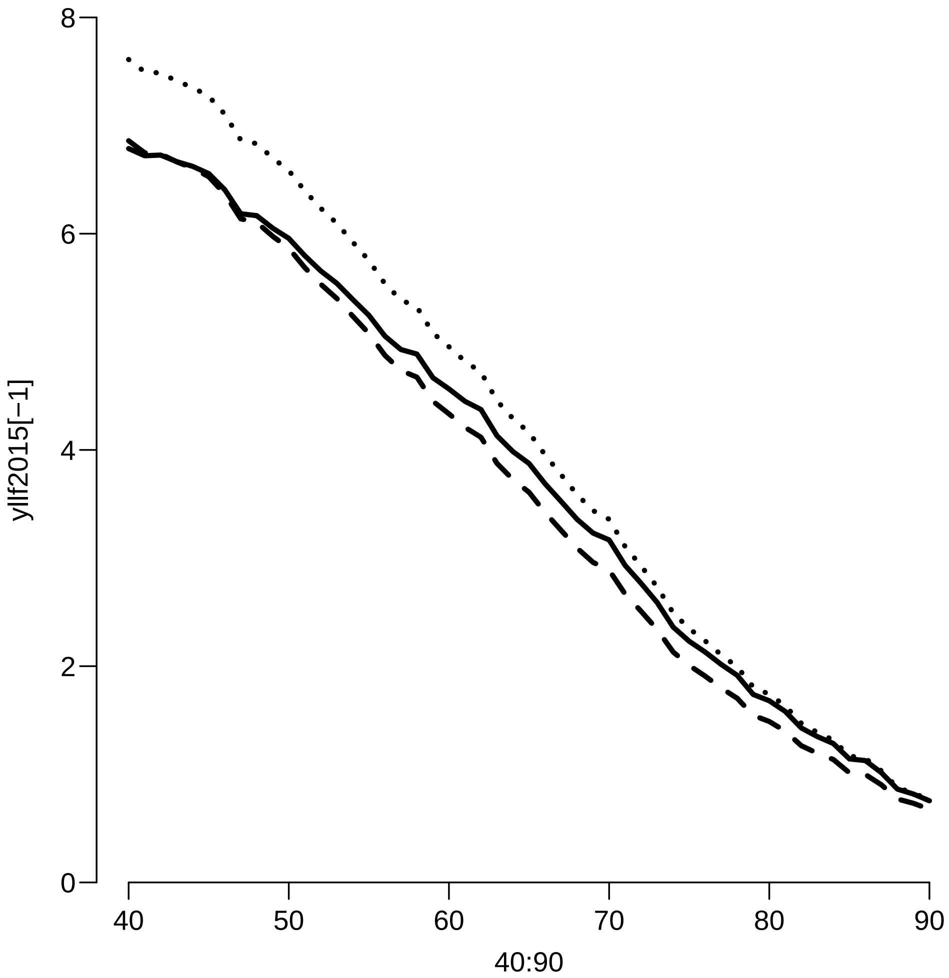

Figure 2.5: *Years of life lost to diabetes in women, using empirical mortality rates from 2015. Full line: Estimates from the illness-death model, dotted line: estimates assuming immunity among non-DM persons, dashed line: approximation using the total population mortality for comparison.*

So from figure 2.5 we see that the immune approach *over*estimates the YLL, mostly in young ages, whereas the approach comparing to the total mortality *under*estimates the YLL, primarily in older ages.

2.4 Modeling / smoothing rates

This part of the exercise replaces the *empirical* rates which are only available in the same form as data (that is 1 × 1 year age by period classes) with *predicted* rates based on a statistical model. Another way of putting this is that we use statistical modeling to smooth the rates, so that they are biologically plausible by having the property that they vary smoothly by age and calendar time.

Moreover, this part of the exercises also collects estimated values of YLL for different combinations of sex, age and calendar time in an *array*, so it is easier to print and plot results.

11. The curves of mortality and just shown are just for women and just for one particular year — cross-sectional mortality rates. This also gives wiggly curves which is just attributable to random error. To some extent this carries over to the YLL curves as well, less so because YLL is a *cumulative* measure.

    We do have a dataset covering the 20-year period 1996–2015 incl. and both for men and women; it would therefore be more timely fit proper statistical models for the mortality and incidence rates over the period 1996–2015 for all ages 0–99, and use predictions from these as input to the functions computing years of life lost.

    The models we use will be age-period-cohort models [1] providing estimated mortality rates for ages 0–99 and dates 1.1.1996–31.12.2015. Since we will use parametric models using age and calendar time as *quantitative* variable we can make predictions for any combination of age and calendar time — in principle also *outside* the period where data are available, of course provided that we are willing to believe the projections of the rates.

    First we transform the age and period variables to reflect the mean age and period in each of the cells in the age by period Lexis diagram, and we compute the total number of deaths and person-years. We also restrict data to ages over 30:

    ```
    > data( DMepi )
    > DMepi <- transform( DMepi, A = A + 0.5,
    +                            P = P + 0.5,
    +                          D.T = D.nD + D.DM,
    +                          Y.T = Y.nD + Y.DM )
    > DMepi <- subset( DMepi, A>30 )
    ```

    With the correct age and period coding of the age and date, we can then fit models for mortality and incidence rates. Note that we for comparative purposes also fit a model for the *total* mortality.

    We will use natural splines (the function `Ns`), and hence we must specify a set of knots for splines for the age, period and cohort effects — we use separate sets for the age-effects in mortality and incidence:

    ```
    > ( a.kn <- seq(40,95,,7) )
    ```

    ```
    [1] 40.00000 49.16667 58.33333 67.50000 76.66667 85.83333 95.00000
    ```

    ```
    > ( i.kn <- seq(35,85,,9) )
    ```

```
[1] 35.00 41.25 47.50 53.75 60.00 66.25 72.50 78.75 85.00
```

```
> ( p.kn <- seq(1997,2015,,5) )
```

```
[1] 1997.0 2001.5 2006.0 2010.5 2015.0
```

```
> ( c.kn <- seq(1910,1975,,7) )
```

```
[1] 1910.000 1920.833 1931.667 1942.500 1953.333 1964.167 1975.000
```

Once we defined the knots, we can check the number of different types of events between knots:

```
> ae <- xtabs( cbind(D.T,D.nD,D.DM,X) ~ cut(A,c(30,a.kn,Inf)) + sex, data=DMepi )
> ftable( addmargins(ae,1), col.vars=3:2 )
```

```
                                D.T          D.nD          D.DM             X       
                         sex      M      F      M      F      M      F      M      F
cut(A, c(30, a.kn, Inf))                                                            
(30,40]                        8642   4287   8165   4057    477    230   7872   6125
(40,49.2]                     18307  10898  16597  10007   1710    891  22976  16695
(49.2,58.3]                   41358  26273  35931  23709   5427   2564  38460  24456
(58.3,67.5]                   86946  58330  71420  50860  15526   7470  52510  35467
(67.5,76.7]                  123844  95912 100813  81764  23031  14148  34745  29303
(76.7,85.8]                  155566 161873 130564 139518  25002  22355  16836  19403
(85.8,95]                     91967 165741  80626 147904  11341  17837   3568   6160
(95,Inf]                       9797  32147   8969  29868    828   2279    150    380
Sum                          536427 555461 453085 487687  83342  67774 177117 137989
```

```
> ie <- xtabs( cbind(D.T,D.nD,D.DM,X) ~ cut(A,c(30,i.kn,Inf)) + sex, data=DMepi )
> ftable( addmargins(ie,1), col.vars=3:2 )
```

```
                                D.T          D.nD          D.DM             X       
                         sex      M      F      M      F      M      F      M      F
cut(A, c(30, i.kn, Inf))                                                            
(30,35]                        3473   1568   3319   1476    154     92   2974   2653
(35,41.2]                      6587   3479   6154   3287    433    192   7428   6447
(41.2,47.5]                   14004   8326  12731   7634   1273    692  17215  11716
(47.5,53.8]                   22380  13932  19684  12701   2696   1231  22733  14516
(53.8,60]                     35402  22696  30227  20241   5175   2455  29178  18519
(60,66.2]                     51911  34732  42720  30307   9191   4425  32412  21622
(66.2,72.5]                   84278  61401  68097  52608  16181   8793  31227  24338
(72.5,78.8]                   95308  79520  78218  67626  17090  11894  18670  17532
(78.8,85]                    105069 111007  88437  95760  16632  15247  10480  12556
(85,Inf]                     118015 218800 103498 196047  14517  22753   4800   8090
Sum                          536427 555461 453085 487687  83342  67774 177117 137989
```

```
> pe <- xtabs( cbind(D.T,D.nD,D.DM,X) ~ cut(P,c(1990,p.kn,Inf)) + sex, data=DMepi )
> ftable( addmargins(pe,1), col.vars=3:2 )
```

```
                                  D.T          D.nD          D.DM             X       
                           sex      M      F      M      F      M      F      M      F
cut(P, c(1990, p.kn, Inf))                                                            
(1990,1997]                      29292  29795  26337  27139   2955   2656   6294   5096
(1997,2002]                     140143 146999 123654 132636  16489  14363  34270  27040
(2002,2006]                     108153 113870  92741 101093  15412  12777  33116  25799
(2006,2010]                     131758 137025 109604 118982  22154  18043  51235  39492
(2010,2015]                     101500 102343  80691  86678  20809  15665  42722  33280
(2015,Inf]                       25581  25429  20058  21159   5523   4270   9480   7282
Sum                             536427 555461 453085 487687  83342  67774 177117 137989
```

```
> ce <- xtabs( cbind(D.T,D.nD,D.DM,X) ~ cut(P-A,c(-Inf,c.kn,Inf)) + sex, data=DMepi )
> ftable( addmargins(ce,1), col.vars=3:2 )
```

```
                                      D.T           D.nD          D.DM                X
                               sex      M      F      M      F      M      F      M
cut(P - A, c(-Inf, c.kn, Inf))
(-Inf,1.91e+03]                     21696  53528  19912  49797   1784   3731    536
(1.91e+03,1.92e+03]                100947 156279  89351 139871  11596  16408   5730
(1.92e+03,1.93e+03]                172751 178116 145838 153694  26913  24422  22062  2
(1.93e+03,1.94e+03]                120506  92151  96478  78028  24028  14123  42920  3
(1.94e+03,1.95e+03]                 77803  50607  63705  44033  14098   6574  56002  3
(1.95e+03,1.96e+03]                 31242  18697  27224  16676   4018   2021  33972  2
(1.96e+03,1.98e+03]                 10102   5387   9268   4935    834    452  14341  1
(1.98e+03, Inf]                      1380    696   1309    653     71     43   1554
Sum                                536427 555461 453085 487687  83342  67774 177117 13
```

Once we have fixed the knots for the age, period and cohort effects we fit Separate APC-models for mortality among non-diseased (`mW`), mortality among diseased (`mD`), total mortality (`mT`) and incidence rates (`iW`), separately for men (suffix `.m`) and for men (suffix `.f`).

```
> mW.m <- glm( D.nD ~ Ns(A   ,knots=a.kn,int=TRUE) +
+                     Ns(  P,knots=p.kn,ref=2005) +
+                     Ns(P-A,knots=c.kn,ref=1950),
+             offset = log(Y.nD),
+             family = poisson,
+               data = subset( DMepi, sex=="M" ) )
> mD.m <- update( mW.m,  D.DM ~ . , offset=log(Y.DM) )
> mT.m <- update( mW.m,  D.T  ~ . , offset=log(Y.T ) )
> iW.m <- glm(    X ~ Ns(A   ,knots=i.kn,int=TRUE) +
+                     Ns(  P,knots=p.kn,ref=2005) +
+                     Ns(P-A,knots=c.kn,ref=1950),
+             offset = log(Y.nD),
+             family = poisson,
+               data = subset( DMepi, sex=="M" ) )
```

The models for women are fitted by simply updating the models for men, using a different subset of the dataset:

```
> mW.f <- update( mW.m, data = subset( DMepi, sex=="F" ) )
> mD.f <- update( mD.m, data = subset( DMepi, sex=="F" ) )
> mT.f <- update( mT.m, data = subset( DMepi, sex=="F" ) )
> iW.f <- update( iW.m, data = subset( DMepi, sex=="F" ) )
```

For comparison we can show how the smoothed rates for the midpoint of 2015 compared to the empirical rates for women for the period of 2015. So we first predict the rates from the models:

```
> nd <- data.frame( A = 30:99+0.5,
+                   P = 2015.5,
+                Y.nD = 1,
+                Y.DM = 1,
+                Y.T  = 1 )
> muW.f <- ci.pred( mW.f, nd )
> muD.f <- ci.pred( mD.f, nd )
> muT.f <- ci.pred( mT.f, nd )
> lam.f <- ci.pred( iW.f, nd )
> clr <- c("red","blue","limegreen","black")
> with( w15, matplot( A, cbind( mW, mD, mT, iW)*1000, lwd=3,
+                      log="y", type="l", lty=1,
+                      col=clr ) )
> matlines( nd$A, cbind(muW.f,muD.f,muT.f,lam.f)*1000,
+           lty=1, lwd=c(3,1,1),
+           col=rep(clr,each=3) )
> cn <- par("usr")
> text( rep( cn[1:2]%*%c(8,2)/10, 4 ),
+       10^((cbind(9:6,1:4)/10)%*%cn[4:3])[c(3,1,2,4)],
+       c("no DM mort","DM mort","Total mort","DM inc"), adj=1,
+       col=clr )
```

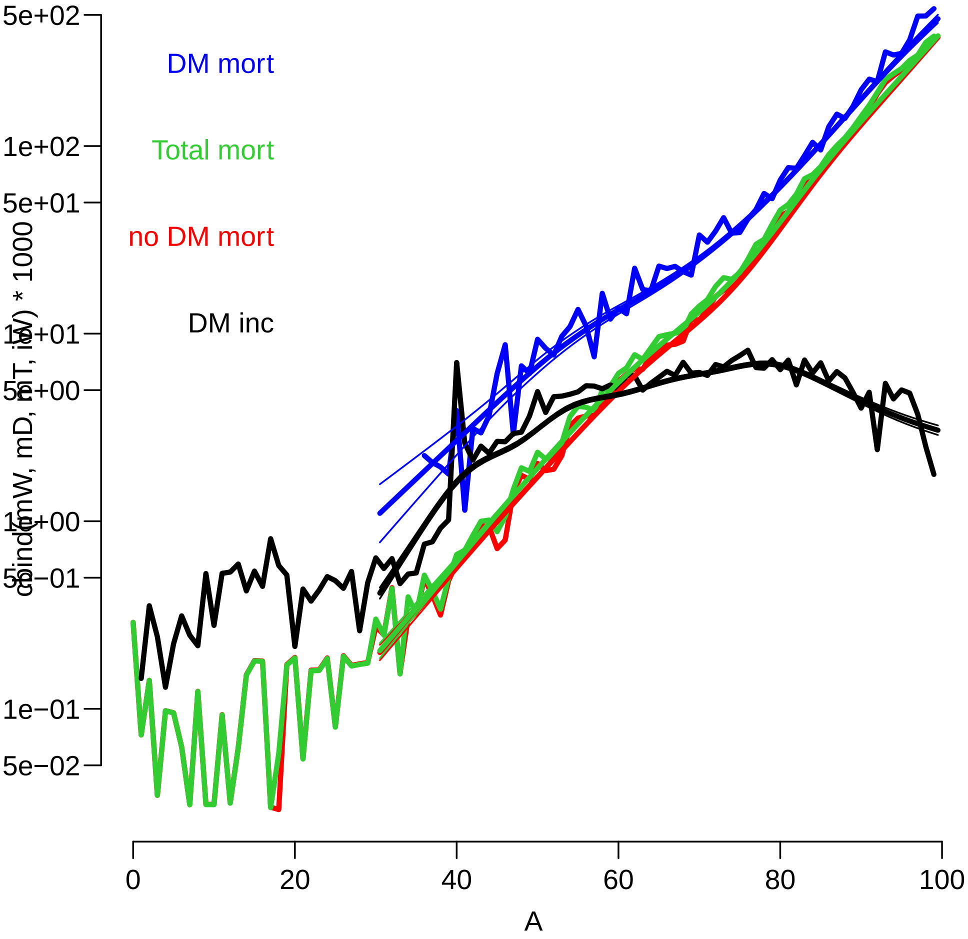

Figure 2.6: *Empirical rates for the year 2015 and fitted rates for women at mid-2015.*

From figure 2.6 we see a much more credible version of the rates; a basis we should use for predicting the years of life lost.

12. Also we can redo the YLL analyses with the smoothed rates, and plot it on top of the empirical rate based ones; first we re-compute and save the YLL based on the empirical rates:

```
> yllf2015s <- with( w15, yll( int=1, muW=mW, muD=mD, lam=iW,
+                                                   age.in=0, A=c(40:90) ) )
> yllf2015i <- with( w15, yll( int=1, muW=mW, muD=mD, age.in=0, A=c(40:90) ) )

NOTE: Calculations assume that Well persons cannot get Ill (quite silly!).

> yllf2015t <- with( w15, yll( int=1, muW=mT, muD=mD, age.in=0, A=c(40:90), note=F ) )
```

and then the corresponding YLL computed from the rates as predicted from the statistical model:

```
> yllsf2015s <- yll( int=1, muW=muW.f[,1], muD=muD.f[,1], lam=lam.f[,1], age.in=30, A=
> yllsf2015i <- yll( int=1, muW=muW.f[,1], muD=muD.f[,1], age.in=30, A=c(40:90) )

NOTE: Calculations assume that Well persons cannot get Ill (quite silly!).

> yllsf2015t <- yll( int=1, muW=muT.f[,1], muD=muD.f[,1], age.in=30, A=c(40:90), note=
```

Finally we can plot these 6 curves together; the wiggly, empirical based as you saw before in gray, and the new model-based in black

```
>  plot( 40:90, yllf2015s [-1], type="l", lwd=3, ylim=c(0,8), yaxs="i", col=gray(0.6)
> lines( 40:90, yllf2015i [-1], type="l", lwd=3, lty="12", col=gray(0.6) )
> lines( 40:90, yllf2015t [-1], type="l", lwd=3, lty="53", col=gray(0.6) )
> # smoothed estimates
> lines( 40:90, yllsf2015s[-1], type="l", lwd=3 )
> lines( 40:90, yllsf2015i[-1], type="l", lwd=3, lty="12" )
> lines( 40:90, yllsf2015t[-1], type="l", lwd=3, lty="53" )
```

13. But we would like to see YLL for different (well, *all*) combinations of sex, age, calendar time and would like to explore the different ways of computing YLL. Hence we set up a 4-dimensional array to collect the YLL as computed by different methods and differences between them by sex, age and date:

```
> # ages and dates for computing YLL
> a.ref <- 30:90
> p.ref <- 1996:2016
> # array to hold results
> aYLL <- NArray( list( type = c("Sus","Imm","Tot","I-S","I-S Ex%","T-S","T-S Ex%"),
+                        sex = levels( DMepi$sex ),
+                        age = a.ref,
+                       date = p.ref ) )
> str( aYLL )

 logi [1:7, 1:2, 1:61, 1:21] NA NA NA NA NA NA ...
 - attr(*, "dimnames")=List of 4
  ..$ type: chr [1:7] "Sus" "Imm" "Tot" "I-S" ...
  ..$ sex : chr [1:2] "M" "F"
  ..$ age : chr [1:61] "30" "31" "32" "33" ...
  ..$ date: chr [1:21] "1996" "1997" "1998" "1999" ...

> dim( aYLL ) ; prod( dim( aYLL ) )

type  sex  age date
   7    2   61   21

[1] 17934
```

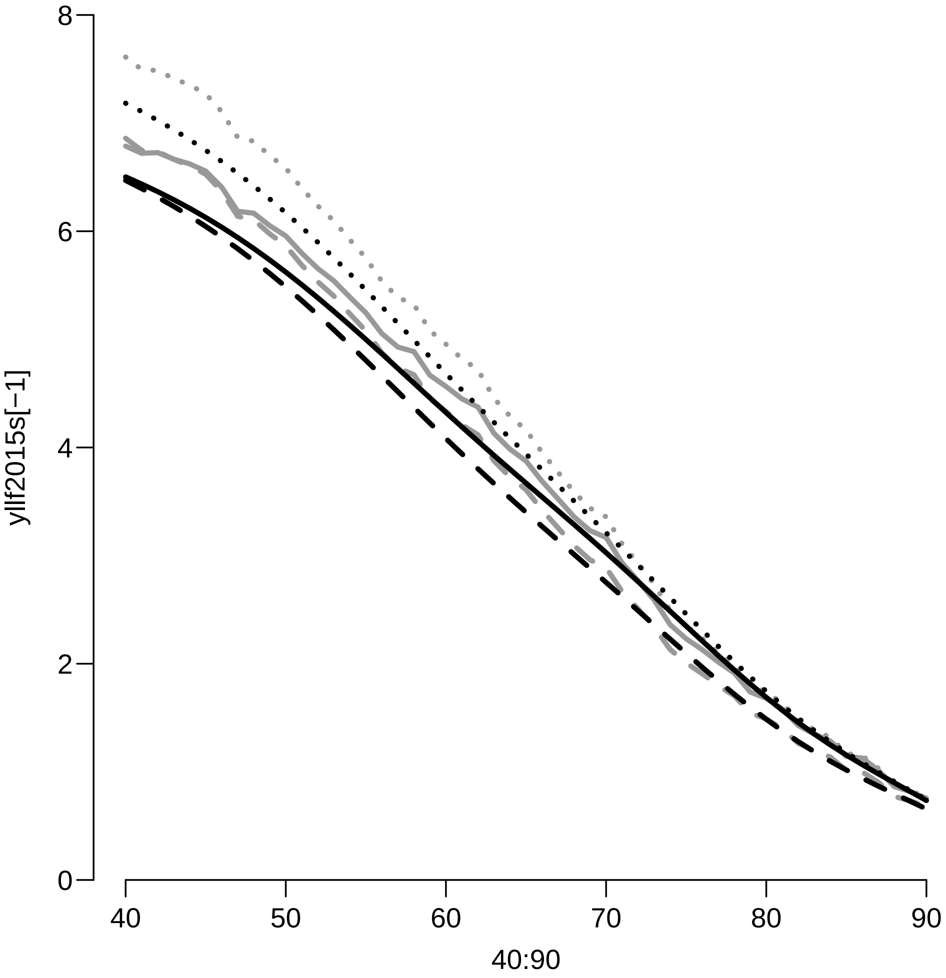

Figure 2.7: *Years of life lost to diabetes in women, using empirical mortality rates from 2015 (gray) and smoothed estimates of rates (black). Full line: Estimates from the illness-death model, dotted line: estimates assuming immunity among non-DM persons, dashed line: approximation using the total population mortality for comparison.*

So `aYLL` is now a 4-dimensional array with a total of 17,934 entries, each representing a YLL fora particular combination of sex, age etc. A particular element (number) in the array can be referenced as, say, `aYLL["Imm","M","35","2001"]`. Note that the `"35"` refer to the entry for 35 year old (the label is `"35"`). We could instead of the names refer to the *number* along each dimension, so the element `aYLL["Imm","M","35","2002"]` is the same as `aYLL[2,1,6,7]` or `aYLL["Imm",1,"35",7]`, since `"35"` is the 6th element in the age-dimension (the 3rd dimension). If we want *all* elements along the age-dimension we just leave the space between commas blank: `aYLL["Imm",1■7]`. But any mention of a subset of the array must have three commas between the square brackets.

Once we have the array, we can fill values in it using a loops over calendar time; the age-dimension is an argument to `yll` and sex is handled in completely different models, as is the three different approaches to computing the YLL:

```
> system.time(
+ for( ip in p.ref )
+    {
+    nd <- data.frame( A = seq(30.2,100,0.2)-0.1,
+                      P = ip,
+                   Y.nD = 1,
+                   Y.DM = 1,
+                   Y.T  = 1 )
+    muW.m <- ci.pred( mW.m, nd )[,1]
+    muD.m <- ci.pred( mD.m, nd )[,1]
+    muT.m <- ci.pred( mT.m, nd )[,1]
+    lam.m <- ci.pred( iW.m, nd )[,1]
+    muW.f <- ci.pred( mW.f, nd )[,1]
+    muD.f <- ci.pred( mD.f, nd )[,1]
+    muT.f <- ci.pred( mT.f, nd )[,1]
+    lam.f <- ci.pred( iW.f, nd )[,1]
+ aYLL["Imm","M",,paste(ip)] <- yll( int=0.2, muW.m, muD.m, lam=NULL , A=a.ref, age.in
+ aYLL["Imm","F",,paste(ip)] <- yll( int=0.2, muW.f, muD.f, lam=NULL , A=a.ref, age.in
+ aYLL["Tot","M",,paste(ip)] <- yll( int=0.2, muT.m, muD.m, lam=NULL , A=a.ref, age.in
+ aYLL["Tot","F",,paste(ip)] <- yll( int=0.2, muT.f, muD.f, lam=NULL , A=a.ref, age.in
+ aYLL["Sus","M",,paste(ip)] <- yll( int=0.2, muW.m, muD.m, lam=lam.m, A=a.ref, age.in
+ aYLL["Sus","F",,paste(ip)] <- yll( int=0.2, muW.f, muD.f, lam=lam.f, A=a.ref, age.in
+    } )

   user  system elapsed
 18.639   0.000  18.638
```

One advantage of an array is that we can perform simple operations over all levels of some of the dimensions:

```
> aYLL["I-S"    ,,,] <-  aYLL["Imm",,,] - aYLL["Sus",,,]
> aYLL["I-S Ex%",,,] <- (aYLL["Imm",,,] - aYLL["Sus",,,]) / aYLL["Sus",,,] * 100
> aYLL["T-S"    ,,,] <-  aYLL["Tot",,,] - aYLL["Sus",,,]
> aYLL["T-S Ex%",,,] <- (aYLL["Tot",,,] - aYLL["Sus",,,]) / aYLL["Sus",,,] * 100
```

And a further advantage is the possibility to get a simple overview of select parts of the values using `ftable`; here we take every 10th age and every 5th date element and list in two different ways, by changing `row.vars` and `col.vars`:

```
> round( ftable( aYLL[,,seq(11,51,10),seq(1,21,5)], col.vars=c(3,2), row.vars=c(4,1) ), 1 )

             age    40          50          60          70          80      
             sex     M     F     M     F     M     F     M     F     M     F
date type                                                                   
1996 Sus           9.3   9.9   7.4   8.5   5.5   6.8   3.6   4.7   1.9   2.6
     Imm          10.2  10.7   8.1   9.2   5.9   7.2   3.8   5.0   2.0   2.7
     Tot           9.6  10.2   7.6   8.7   5.5   6.8   3.5   4.6   1.8   2.5
     I-S           0.9   0.8   0.7   0.7   0.5   0.5   0.2   0.3   0.1   0.1
     I-S Ex%       9.6   7.9   9.8   7.9   8.6   7.1   6.1   5.4   3.7   3.2
     T-S           0.3   0.3   0.2   0.2   0.0   0.0  -0.1  -0.1  -0.1  -0.1
     T-S Ex%       3.6   2.8   2.7   2.1   0.5   0.4  -2.6  -2.2  -4.6  -4.3
2001 Sus           8.6   8.4   6.9   7.2   5.2   5.7   3.5   4.0   2.0   2.3
     Imm           9.7   9.2   7.8   7.9   5.8   6.2   3.8   4.3   2.0   2.4
     Tot           9.0   8.7   7.1   7.3   5.2   5.7   3.4   3.9   1.8   2.2
     I-S           1.1   0.8   0.9   0.7   0.6   0.5   0.3   0.3   0.1   0.1
     I-S Ex%      12.2   9.6  12.5   9.5  10.8   8.6   7.5   6.4   4.2   3.5
     T-S           0.3   0.3   0.2   0.2   0.0   0.0  -0.1  -0.1  -0.1  -0.1
     T-S Ex%       3.8   3.2   2.6   2.3  -0.2   0.4  -3.9  -2.6  -6.3  -5.2
```

```
2006 Sus           7.5   7.6   6.0   6.4   4.6   5.1   3.1   3.6   1.7   2.1
     Imm           8.6   8.4   6.9   7.1   5.2   5.6   3.3   3.8   1.8   2.2
     Tot           7.8   7.8   6.2   6.5   4.5   5.1   2.9   3.4   1.6   2.0
     I-S           1.1   0.9   0.9   0.7   0.6   0.5   0.3   0.3   0.1   0.1
     I-S Ex%      14.2  11.3  14.4  11.0  12.6   9.8   8.6   7.5   4.6   4.0
     T-S           0.3   0.2   0.1   0.1  -0.1   0.0  -0.2  -0.1  -0.1  -0.1
     T-S Ex%       3.7   3.3   2.1   2.0  -1.3  -0.2  -5.5  -3.5  -7.7  -6.4
2011 Sus           6.5   6.5   5.3   5.5   4.1   4.3   2.8   3.0   1.5   1.8
     Imm           7.8   7.5   6.3   6.3   4.7   4.8   3.1   3.3   1.6   1.8
     Tot           6.9   6.8   5.5   5.7   4.0   4.3   2.6   2.9   1.4   1.6
     I-S           1.3   1.0   1.0   0.8   0.7   0.5   0.3   0.3   0.1   0.1
     I-S Ex%      19.5  15.1  19.7  14.4  17.2  12.7  11.8   9.6   6.1   5.2
     T-S           0.4   0.3   0.2   0.2   0.0   0.0  -0.2  -0.1  -0.2  -0.1
     T-S Ex%       5.7   5.1   3.6   3.2  -1.1   0.1  -6.9  -4.1 -10.3  -7.8
2016 Sus           6.5   6.5   5.3   5.7   4.1   4.4   2.9   3.1   1.6   1.7
     Imm           7.3   7.2   6.1   6.2   4.6   4.7   3.1   3.2   1.7   1.8
     Tot           6.3   6.5   5.1   5.5   3.7   4.1   2.5   2.8   1.3   1.5
     I-S           0.9   0.7   0.7   0.5   0.5   0.4   0.2   0.2   0.1   0.1
     I-S Ex%      13.4  10.1  13.5   9.4  11.7   8.1   8.1   6.0   4.1   3.3
     T-S          -0.1  -0.1  -0.2  -0.2  -0.4  -0.3  -0.4  -0.3  -0.2  -0.2
     T-S Ex%      -1.9  -1.0  -4.3  -2.9  -8.8  -6.0 -13.4  -9.4 -15.7 -12.2
```

```
> round( ftable( aYLL[,,seq(11,51,10),seq(1,21,5)], col.vars=c(2,3), row.vars=c(1,4) )
```

```
             sex     M                             F
             age    40    50    60    70    80    40    50    60    70    80
type    date
Sus     1996       9.3   7.4   5.5   3.6   1.9   9.9   8.5   6.8   4.7   2.6
        2001       8.6   6.9   5.2   3.5   2.0   8.4   7.2   5.7   4.0   2.3
        2006       7.5   6.0   4.6   3.1   1.7   7.6   6.4   5.1   3.6   2.1
        2011       6.5   5.3   4.1   2.8   1.5   6.5   5.5   4.3   3.0   1.8
        2016       6.5   5.3   4.1   2.9   1.6   6.5   5.7   4.4   3.1   1.7
Imm     1996      10.2   8.1   5.9   3.8   2.0  10.7   9.2   7.2   5.0   2.7
        2001       9.7   7.8   5.8   3.8   2.0   9.2   7.9   6.2   4.3   2.4
        2006       8.6   6.9   5.2   3.3   1.8   8.4   7.1   5.6   3.8   2.2
        2011       7.8   6.3   4.7   3.1   1.6   7.5   6.3   4.8   3.3   1.8
        2016       7.3   6.1   4.6   3.1   1.7   7.2   6.2   4.7   3.2   1.8
Tot     1996       9.6   7.6   5.5   3.5   1.8  10.2   8.7   6.8   4.6   2.5
        2001       9.0   7.1   5.2   3.4   1.8   8.7   7.3   5.7   3.9   2.2
        2006       7.8   6.2   4.5   2.9   1.6   7.8   6.5   5.1   3.4   2.0
        2011       6.9   5.5   4.0   2.6   1.4   6.8   5.7   4.3   2.9   1.6
        2016       6.3   5.1   3.7   2.5   1.3   6.5   5.5   4.1   2.8   1.5
I-S     1996       0.9   0.7   0.5   0.2   0.1   0.8   0.7   0.5   0.3   0.1
        2001       1.1   0.9   0.6   0.3   0.1   0.8   0.7   0.5   0.3   0.1
        2006       1.1   0.9   0.6   0.3   0.1   0.9   0.7   0.5   0.3   0.1
        2011       1.3   1.0   0.7   0.3   0.1   1.0   0.8   0.5   0.3   0.1
        2016       0.9   0.7   0.5   0.2   0.1   0.7   0.5   0.4   0.2   0.1
I-S Ex% 1996       9.6   9.8   8.6   6.1   3.7   7.9   7.9   7.1   5.4   3.2
        2001      12.2  12.5  10.8   7.5   4.2   9.6   9.5   8.6   6.4   3.5
        2006      14.2  14.4  12.6   8.6   4.6  11.3  11.0   9.8   7.5   4.0
        2011      19.5  19.7  17.2  11.8   6.1  15.1  14.4  12.7   9.6   5.2
        2016      13.4  13.5  11.7   8.1   4.1  10.1   9.4   8.1   6.0   3.3
T-S     1996       0.3   0.2   0.0  -0.1  -0.1   0.3   0.2   0.0  -0.1  -0.1
        2001       0.3   0.2   0.0  -0.1  -0.1   0.3   0.2   0.0  -0.1  -0.1
        2006       0.3   0.1  -0.1  -0.2  -0.1   0.2   0.1   0.0  -0.1  -0.1
        2011       0.4   0.2   0.0  -0.2  -0.2   0.3   0.2   0.0  -0.1  -0.1
        2016      -0.1  -0.2  -0.4  -0.4  -0.2  -0.1  -0.2  -0.3  -0.3  -0.2
T-S Ex% 1996       3.6   2.7   0.5  -2.6  -4.6   2.8   2.1   0.4  -2.2  -4.3
        2001       3.8   2.6  -0.2  -3.9  -6.3   3.2   2.3   0.4  -2.6  -5.2
```

```
2006        3.7   2.1  -1.3  -5.5  -7.7   3.3   2.0  -0.2  -3.5  -6.4
2011        5.7   3.6  -1.1  -6.9 -10.3   5.1   3.2   0.1  -4.1  -7.8
2016       -1.9  -4.3  -8.8 -13.4 -15.7  -1.0  -2.9  -6.0  -9.4 -12.2
```

14. Plotting the age- and time trends in the YLL is done by selecting relevant parts of the array; for example `aYLL["Sus","M",,]` is an age by period matrix of years of life lost to DM for men, using the correct model (susceptible).

    So we plot this family of curves showing the change over time, one panel for men and one panel for women

    ```
    > par( mfrow=c(1,2), mar=c(3,3,1,1), mgp=c(3,1,0)/1.6, bty="n", las=1 )
    > matplot( a.ref, aYLL["Sus","M",,],
    +          type="l", lty=1, col="blue", lwd=1:2,
    +          ylim=c(0,12), xlab="Age",
    +          ylab="Years lost to DM", yaxs="i" )
    > abline(v=50,h=1:10,col=gray(0.7))
    > text( 90, 11, "Men", col="blue", adj=1 )
    > text( 40, aYLL["Sus","M","40","1996"], "1996", adj=c(0,0), col="blue" )
    > text( 43, aYLL["Sus","M","44","2016"], "2016", adj=c(1,1), col="blue" )
    > matplot( a.ref, aYLL["Sus","F",,],
    +          type="l", lty=1, col="red", lwd=1:2,
    +          ylim=c(0,12), xlab="Age",
    +          ylab="Years lost to DM", yaxs="i" )
    > abline(v=50,h=1:10,col=gray(0.7))
    > text( 90, 11, "Women", col="red", adj=1 )
    > text( 40, aYLL["Sus","F","40","1996"], "1996", adj=c(0,0), col="red" )
    > text( 43, aYLL["Sus","F","44","2016"], "2016", adj=c(1,1), col="red" )
    ```

15. A similar piece of code is needed to make the corresponding curves for `Imm` and `Tot`, so we simply put it all in a function where the type of calculation is the argument. All other variables will just be taken from the global environment, so the function defined here will only work at this place in the program.

    ```
    > plyll <- function(wh){
    + par( mfrow=c(1,2), mar=c(3,3,1,1), mgp=c(3,1,0)/1.6, bty="n", las=1 )
    +
    + matplot( a.ref, aYLL[wh,"M",,],
    +          type="l", lty=1, col="blue", lwd=1:2,
    +          ylim=c(0,12), xlab="Age",
    +          ylab="Years lost to DM", yaxs="i" )
    + abline(v=50,h=1:12,col=gray(0.7))
    + text( 90, 11.5, "Men", col="blue", adj=1 )
    + text( 40, aYLL[wh,"M","40","1996"], "1996", adj=c(0,0), col="blue" )
    + text( 43, aYLL[wh,"M","44","2016"], "2016", adj=c(1,1), col="blue" )
    +
    + matplot( a.ref, aYLL[wh,"F",,],
    +          type="l", lty=1, col="red", lwd=1:2,
    +          ylim=c(0,12), xlab="Age",
    +          ylab="Years lost to DM", yaxs="i" )
    + abline(v=50,h=1:12,col=gray(0.7))
    + text( 90, 11.5, "Women", col="red", adj=1 )
    + text( 40, aYLL[wh,"F","40","1996"], "1996", adj=c(0,0), col="red" )
    + text( 43, aYLL[wh,"F","44","2016"], "2016", adj=c(1,1), col="red" )
    + }
    > plyll("Imm")
    ```

```
> plyll("Tot")
```

```
> plyll("Sus")
```

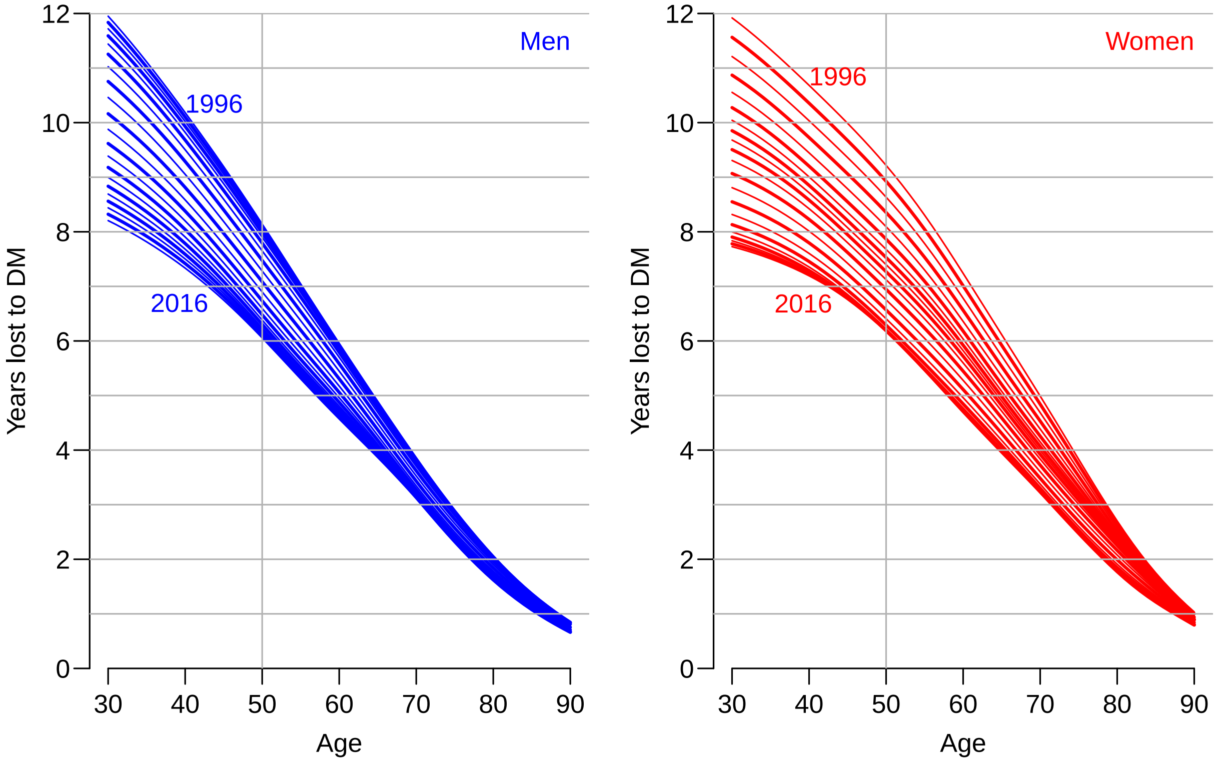

Figure 2.8: *Years of life lost to DM: the difference in expected residual life time at different ages between persons with and without diabetes, assuming the persons without diabetes at a given age remain free from diabetes (immunity assumption). The lines refer to date of evaluation; the top lines refer to 1.1.1996 the bottom ones to 1.1.2012. Blue curves are men, red women.*

From figure 2.9 we see that for men aged 50 the years lost to diabetes has decreased from a bit over 8 to a bit less than 6 years, and for women from 8.5 to 5 years; so a greater improvement for women.

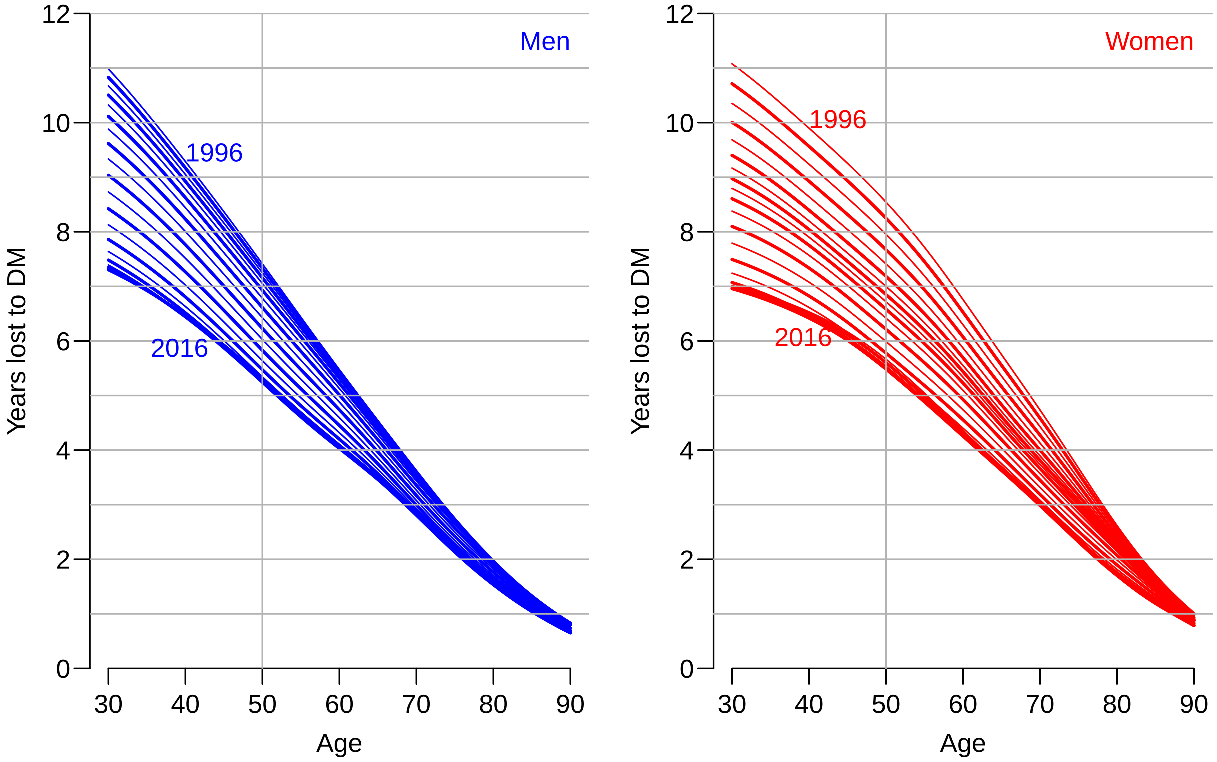

Figure 2.9: *Years of life lost to DM: the difference in expected residual life time at different ages between persons with and without diabetes, allowing the persons without diabetes at a given to contract diabetes and thus be subject to higher mortality. The lines refer to date of evaluation; the top lines refer to 1.1.1996 the bottom ones to 1.1.2012. Blue curves are men, red women.*

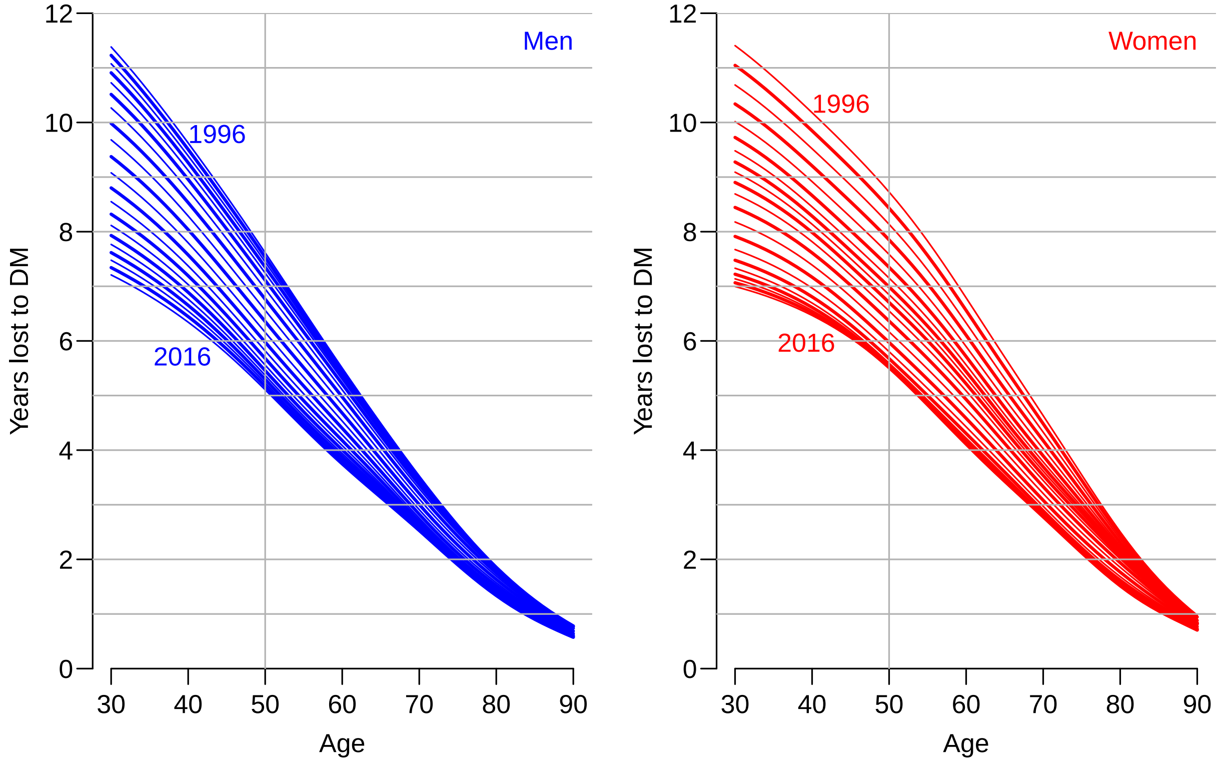

Figure 2.10: *Years of life lost to DM: the difference in expected residual life time at different ages between persons with and without diabetes. Allowance for susceptibility is approximated by using the total population mortality instead of non-DM mortality. The lines refer to date of evaluation; the top lines refer to 1.1.1996 the bottom ones to 1.1.2012. Blue curves are men, red women.*

2.5 Comparing men and women

16. It is illustrative to see the lines for men and women overlaid in the same plot — here men and women overlaid and plots side by side for two different approaches to computation of YLL:

```
> par( mfrow=c(1,2), mar=c(3,3,1,1), mgp=c(3,1,0)/1.6, bty="n", las=1 )
> matplot( a.ref, cbind(aYLL["Imm","M",,],aYLL["Imm","F",,]),
+          type="l", lty=1, col=rep(c("blue","red"),each=18), lwd=1:2,
+          ylim=c(0,12), xlab="Age",
+          ylab="Years lost to DM", yaxs="i" )
> abline(v=50,h=1:12,col=gray(0.7))
> text( 40, aYLL["Imm","F","40","1996"], "1996", adj=c(0,0) )
> text( 43, aYLL["Imm","F","44","2016"], "2016", adj=c(1,1) )
> mtext( "Immunity to DM", side=3 )
> matplot( a.ref, cbind(aYLL["Sus","M",,],aYLL["Sus","F",,]),
+          type="l", lty=1, col=rep(c("blue","red"),each=18), lwd=1:2,
+          ylim=c(0,12), xlab="Age",
+          ylab="Years lost to DM", yaxs="i" )
> abline(v=50,h=1:12,col=gray(0.7))
> text( 40, aYLL["Sus","F","40","1996"], "1996", adj=c(0,0) )
> text( 43, aYLL["Sus","F","44","2016"], "2016", adj=c(1,1) )
> mtext( "Susceptible to DM", side=3 )
```

From figure 2.11 we see that the improvement has been larger for women than for men, but it should be remembered that women have a longer life expectancy then men. Under the (unrealistic) assumption of immunity the improvement in years of life lost to DM for 50-year old women were from 9.2 to 6.2 years and for men from 8.1 to 6.1 years. Under the more realistic assumption that the non-diseased comparison group is allowed to acquire diabetes after the conditioning age, the drop in YLL for women were from 8.5 to 5.7 and for men from 7.4 to 5.3 years.

Also from the tables above we see that the general pattern in the difference between the naive (Immune) and the more realistic (Susceptible) is an overestimation of the years of life lost by about 1 year, or some 5–10%. That is not dramatic, but certainly neither negligible.

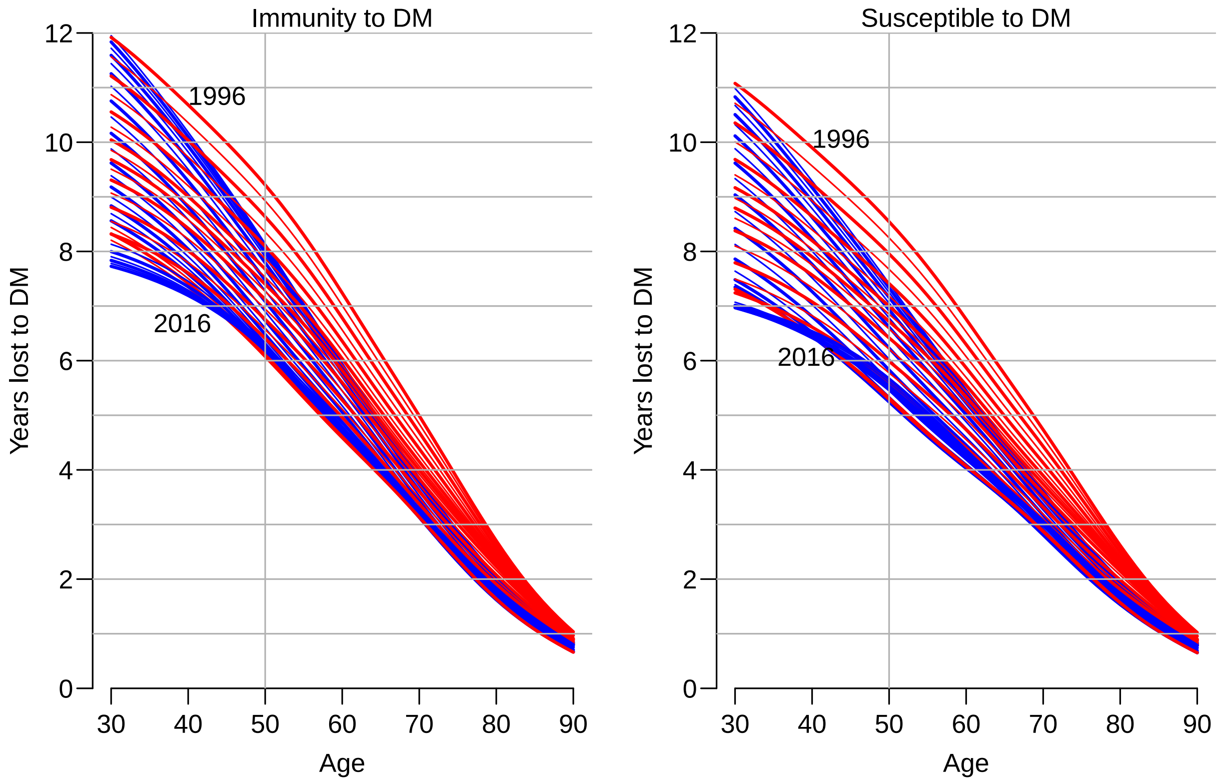

Figure 2.11: *Years of life lost to DM: the difference in expected residual life time at different ages between persons with and without diabetes. The left panel is based on an assumption that persons without DM at a given age will not contract DM; the right panel is based on a model where persons at a given age can contract diabetes and thus transfer to a state with higher mortality. The lines refer to date of evaluation; the top lines refer to 1.1.1996 the bottom ones to 1.1.2016. Blue curves are men, red women.*

2.6 Comparing approaches

17. In order to compare the different approaches to computing years of life lost, we plot the results from the three different approaches in three select years:

```
> cpyll <-
+ function(wh)
+ {
+ par( mfrow=c(1,2), mar=c(3,3,1,1), mgp=c(3,1,0)/1.6, bty="n", las=1 )
+ for( sx in c("M","F") )
+    {
+ matplot( a.ref, t(aYLL[c("Sus","Tot","Imm"),sx,,paste(wh[1])]),
+          col="transparent",
+          ylim=c(0,12), xlab="Age", ylab="Years lost to DM", yaxs="i" )
+ for( yy in wh )
+    matlines( a.ref, t(aYLL[c("Sus","Tot","Imm"),sx,,paste(yy)]),
+              type="l", lty=c("solid","63","22"), lend="butt",
+              col=if(sx=="M") "blue" else "red", lwd=2 )
+ abline(h=1:12,v=50,col=gray(0.7))
+ text( 90, 11.5, sx, col=if( sx=="M") "blue" else "red", adj=1 )
+    }
+ }
> cpyll( seq(1996,2016,10) )
```

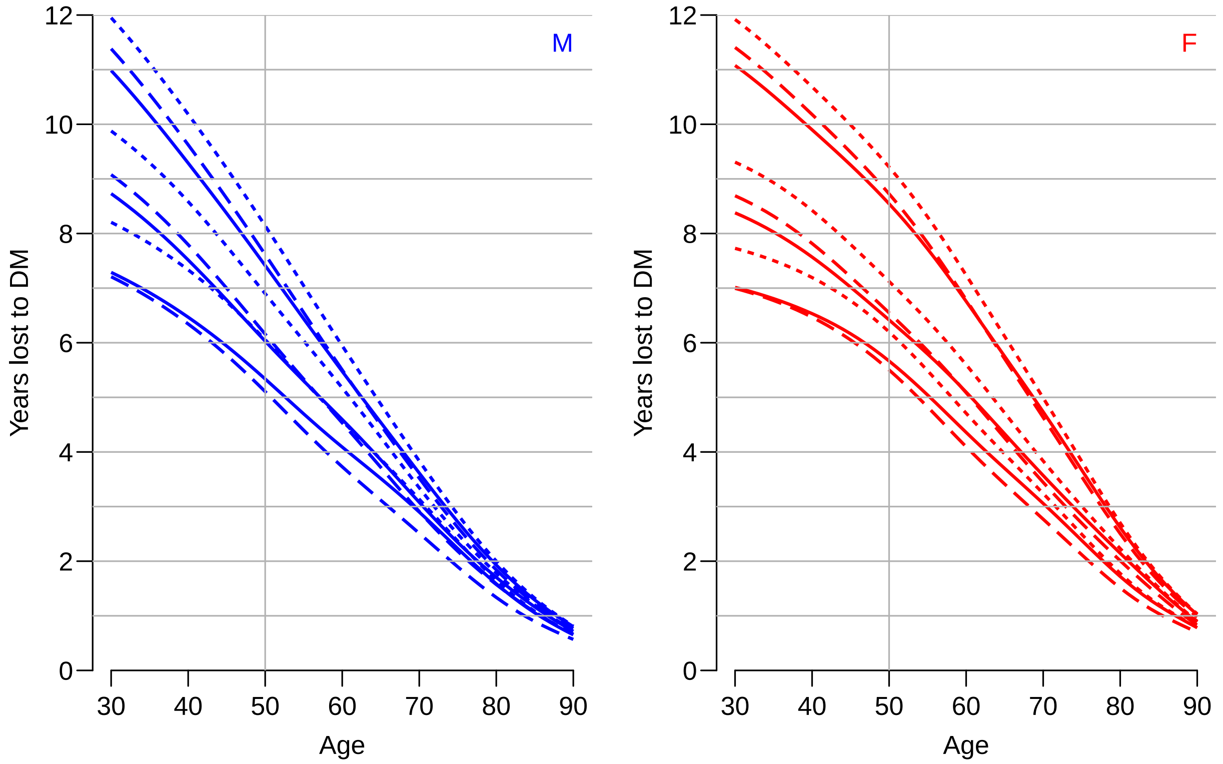

Figure 2.12: *Years of life lost to DM: the difference in expected residual life time at different ages between persons with and without diabetes at the beginning of the years 1996, 2006, and 2016 (top to bottom). Full lines are from using the correct calculation in the illness-death model, dotted lines using the naive approach (immunity assumption) and the broken lines the comparison with the total mortality. Blue curves are men, red women.*

From figure 2.12 we see that the immunity assumption invariably overestimates the years of life lost, at age 50 by something between 6 months and a year, whereas the approximation using the total population mortality (including the diabetics) is much closer to the correct calculation.

The conclusion is that the calculation of YLL should preferably be based on all rates in an illness-death model, and if this is not possible for want of access to diabetes incidence rates, then based on a comparison of the mortality among DM patient and the *total* population mortality.

Chapter 3

Concepts in survival and demography

The following is a summary of relations between various quantities used in analysis of follow-up studies. They are ubiquitous in the analysis and reporting of results. Hence it is important to be familiar with all of them and the relation between them.

3.1 Probability

Survival function:

$$\begin{aligned} S(t) &= \mathrm{P}\{\text{survival at least till } t\} \\ &= \mathrm{P}\{T > t\} = 1 - \mathrm{P}\{T \leq t\} = 1 - F(t) \end{aligned}$$

Conditional survival function:

$$\begin{aligned} S(t|t_{\text{entry}}) &= \mathrm{P}\{\text{survival at least till } t|\text{ alive at } t_{\text{entry}}\} \\ &= S(t)/S(t_{\text{entry}}) \end{aligned}$$

Cumulative distribution function of death times (cumulative risk):

$$\begin{aligned} F(t) &= \mathrm{P}\{\text{death before } t\} \\ &= \mathrm{P}\{T \leq t\} = 1 - S(t) \end{aligned}$$

Density function of death times:

$$f(t) = \lim_{h \to 0} \mathrm{P}\{\text{death in } (t, t+h)\}/h = \lim_{h \to 0} \frac{F(t+h) - F(t)}{h} = F'(t)$$

Intensity:

$$\begin{aligned} \lambda(t) &= \lim_{h \to 0} \mathrm{P}\{\text{event in } (t, t+h] \mid \text{alive at } t\}/h \\ &= \lim_{h \to 0} \frac{F(t+h) - F(t)}{S(t)h} = \frac{f(t)}{S(t)} \\ &= \lim_{h \to 0} -\frac{S(t+h) - S(t)}{S(t)h} = -\frac{\mathrm{d}\log S(t)}{\mathrm{d}t} \end{aligned}$$

The intensity is also known as the hazard function, hazard rate, mortality/morbidity rate or simply “rate”.

Note that f and λ are *scaled* quantities, they have dimension time^{-1}.

Relationships between terms:

$$\begin{aligned} -\frac{\mathrm{d}\log S(t)}{\mathrm{d}t} &= \lambda(t) \\ &\Updownarrow \\ S(t) &= \exp\left(-\int_0^t \lambda(u)\,\mathrm{d}u\right) = \exp\big(-\Lambda(t)\big) \end{aligned}$$

The quantity $\Lambda(t) = \int_0^t \lambda(s)\,\mathrm{d}s$ is called the *integrated intensity* or the **cumulative rate**. It is *not* an intensity (rate), it is dimensionless, despite its name.

$$\lambda(t) = -\frac{\mathrm{d}\log(S(t))}{\mathrm{d}t} = -\frac{S'(t)}{S(t)} = \frac{F'(t)}{1-F(t)} = \frac{f(t)}{S(t)}$$

The cumulative *risk* of an event (to time t) is:

$$F(t) = \mathrm{P}\{\text{Event before time } t\} = \int_0^t \lambda(u)S(u)\,\mathrm{d}u = 1 - S(t) = 1 - \mathrm{e}^{-\Lambda(t)}$$

For small $|x|$ (< 0.05), we have that $1 - \mathrm{e}^{-x} \approx x$, so for small values of the integrated intensity:

$$\text{Cumulative risk to time } t \approx \Lambda(t) = \text{Cumulative rate}$$

3.2 Statistics

Likelihood contribution from follow up of one person:
The likelihood from a number of small pieces of follow-up from one individual is a product of conditional probabilities:

$$\begin{aligned} \mathrm{P}\{\text{event at } t_4|\text{entry at } t_0\} &= \mathrm{P}\{\text{survive } (t_0,t_1)|\text{ alive at } t_0\} \times \\ &\quad\ \mathrm{P}\{\text{survive } (t_1,t_2)|\text{ alive at } t_1\} \times \\ &\quad\ \mathrm{P}\{\text{survive } (t_2,t_3)|\text{ alive at } t_2\} \times \\ &\quad\ \mathrm{P}\{\text{event at } t_4|\text{ alive at } t_3\} \end{aligned}$$

Each term in this expression corresponds to one *empirical rate*[1]
$(d, y) = (\#\text{deaths}, \#\text{risk time})$, i.e. the data obtained from the follow-up of one person in the interval of length y. Each person can contribute many empirical rates, most with $d = 0$; d can only be 1 for the *last* empirical rate for a person.

Log-likelihood for one empirical rate (d, y):

$$\ell(\lambda) = d\log(\lambda) - \lambda y$$

This is under the assumption that the rate (λ) is constant over the interval that the empirical rate refers to.

[1] This is a concept coined by BxC, and so is not necessarily generally recognized.

Log-likelihood for several persons. Adding log-likelihoods from a group of persons (only contributions with identical rates) gives:

$$D\log(\lambda) - \lambda Y,$$

where Y is the total follow-up time, and D is the total number of failures.

Note: The Poisson log-likelihood for an observation D with mean λY is:

$$D\log(\lambda Y) - \lambda Y = D\log(\lambda) + D\log(Y) - \lambda Y$$

The term $D\log(Y)$ does not involve the parameter λ, so the likelihood for an observed rate can be maximized by pretending that the no. of cases D is Poisson with mean λY. But this does *not* imply that D follows a Poisson-distribution. It is entirely a likelihood based computational convenience. Anything that is not likelihood based is not justified.

A linear model for the log-rate, $\log(\lambda) = X\beta$ implies that

$$\lambda Y = \exp\bigl(\log(\lambda) + \log(Y)\bigr) = \exp\bigl(X\beta + \log(Y)\bigr)$$

Therefore, in order to get a linear model for $\log(\lambda)$ we must require that $\log(Y)$ appear as a variable in the model for $D \ \sim (\lambda Y)$ with the regression coefficient fixed to 1, a so-called *offset*-term in the linear predictor.

3.3 Competing risks

Competing risks: If there is more than one, say 3, causes of death, occurring with (cause-specific) rates λ_1, λ_2, λ_3, that is:

$$\lambda_c(a) = \lim_{h\to 0} \mathrm{P}\{\text{death from cause } c \text{ in } (a, a+h] \mid \text{alive at } a\}/h, \quad c = 1, 2, 3$$

The survival function is then:

$$S(a) = \exp\left(-\int_0^a \lambda_1(u) + \lambda_2(u) + \lambda_3(u)\,\mathrm{d}u\right)$$

because you have to escape all 3 causes of death. The probability of dying from cause 1 before age a (the cause-specific cumulative risk) is:

$$\mathrm{P}\{\text{dead from cause 1 at } a\} = \int_0^a \lambda_1(u)S(u)\,\mathrm{d}u \neq 1 - \exp\left(-\int_0^a \lambda_1(u)\,\mathrm{d}u\right)$$

The term $\exp(-\int_0^a \lambda_1(u)\,\mathrm{d}u)$ is sometimes referred to as the “cause-specific survival”, but it does not have any probabilistic interpretation in the real world. It is the survival under the assumption that only cause 1 existed and that the mortality rate from this cause was the same as when the other causes were present too.

Together with the survival function, the cause-specific cumulative risks represent a classification of the population at any time in those alive and those dead from causes 1, 2 and 3 respectively:

$$1 = S(a) + \int_0^a \lambda_1(u)S(u)\,\mathrm{d}u + \int_0^a \lambda_2(u)S(u)\,\mathrm{d}u + \int_0^a \lambda_3(u)S(u)\,\mathrm{d}u, \quad \forall a$$

Subdistribution hazard Fine and Gray defined models for the so-called subdistribution hazard. Recall the relationship between between the hazard (λ) and the cumulative risk (F):

$$\lambda(a) = -\frac{\mathrm{d}\log\bigl(S(a)\bigr)}{\mathrm{d}a} = -\frac{\mathrm{d}\log\bigl(1-F(a)\bigr)}{\mathrm{d}a}$$

When more competing causes of death are present the Fine and Gray idea is to use this transformation to the cause-specific cumulative risk for cause 1, say:

$$\tilde{\lambda}_1(a) = -\frac{\mathrm{d}\log\bigl(1-F_1(a)\bigr)}{\mathrm{d}a}$$

This is what is called the subdistribution hazard, it depends on the survival function S, which depends on *all* the cause-specific hazards:

$$F_1(a) = \mathrm{P}\{\text{dead from cause 1 at } a\} = \int_0^a \lambda_1(u)S(u)\,\mathrm{d}u$$

The subdistribution hazard is merely a transformation of the cause-specific cumulative risk. Namely the same transformation which in the single-cause case transforms the cumulative risk to the hazard.

3.4 Demography

Expected residual lifetime: The expected lifetime (at birth) is simply the variable age (a) integrated with respect to the distribution of age at death:

$$\mathrm{EL} = \int_0^\infty a f(a)\,\mathrm{d}a$$

where f is the density of the distribution of lifetime (age at death).

The relation between the density f and the survival function S is $f(a) = -S'(a)$, so integration by parts gives:

$$\mathrm{EL} = \int_0^\infty a\bigl(-S'(a)\bigr)\,\mathrm{d}a = -\Bigl[aS(a)\Bigr]_0^\infty + \int_0^\infty S(a)\,\mathrm{d}a$$

The first of the resulting terms is 0 because $S(a)$ is 0 at the upper limit and a by definition is 0 at the lower limit.

Hence the expected lifetime can be computed as the integral of the survival function.

The expected *residual* lifetime at age a is calculated as the integral of the *conditional* survival function for a person aged a:

$$\mathrm{EL}(a) = \int_a^\infty S(u)/S(a)\,\mathrm{d}u$$

Lifetime lost due to a disease is the difference between the expected residual lifetime for a diseased person and a non-diseased (well) person at the same age. So all that is needed is a(n estimate of the) survival function in each of the two groups.

$$\mathrm{LL}(a) = \int_a^\infty S_{\text{Well}}(u)/S_{\text{Well}}(a) - S_{\text{Diseased}}(u)/S_{\text{Diseased}}(a)\,\mathrm{d}u$$

Note that the definition of the survival function for a non-diseased person requires a decision as to whether one will consider non-diseased persons immune to the disease in question or not. That is whether we will include the possibility of a well person getting ill and subsequently die. This does not show up in the formulae, but is a decision required in order to devise an estimate of S_{Well}.

Lifetime lost by cause of death is using the fact that the difference between the survival probabilities is the same as the difference between the death probabilities. If several causes of death (3, say) are considered then:

$$\begin{aligned} S(a) = 1 &- \text{P}\{\text{dead from cause 1 at } a\} \\ &- \text{P}\{\text{dead from cause 2 at } a\} \\ &- \text{P}\{\text{dead from cause 3 at } a\} \end{aligned}$$

and hence:

$$\begin{aligned} S_{\text{Well}}(a) - S_{\text{Diseased}}(a) &= \text{P}\{\text{dead from cause 1 at } a|\text{Diseased}\} \\ &+ \text{P}\{\text{dead from cause 2 at } a|\text{Diseased}\} \\ &+ \text{P}\{\text{dead from cause 3 at } a|\text{Diseased}\} \\ &- \text{P}\{\text{dead from cause 1 at } a|\text{Well}\} \\ &- \text{P}\{\text{dead from cause 2 at } a|\text{Well}\} \\ &- \text{P}\{\text{dead from cause 3 at } a|\text{Well}\} \end{aligned}$$

So we can conveniently define the lifetime lost due to cause 2, say, by:

$$\begin{aligned} \text{LL}_2(a) = \int_a^\infty & \text{P}\{\text{dead from cause 2 at } u|\text{Diseased \& alive at } a\} \\ & - \text{P}\{\text{dead from cause 2 at } u|\text{Well \& alive at } a\} \,\text{d}u \end{aligned}$$

These quantities have the property that their sum is the total years of life lost due to the disease:

$$\text{LL}(a) = \text{LL}_1(a) + \text{LL}_2(a) + \text{LL}_3(a)$$

The terms in the integral are computed as (see the section on competing risks):

$$\begin{aligned} \text{P}\{\text{dead from cause 2 at } x|\text{Diseased \& alive at } a\} &= \int_a^x \lambda_{2,\text{Dis}}(u) S_{\text{Dis}}(u)/S_{\text{Dis}}(a) \,\text{d}u \\ \text{P}\{\text{dead from cause 2 at } x|\text{Well \& alive at } a\} &= \int_a^x \lambda_{2,\text{Well}}(u) S_{\text{Well}}(u)/S_{\text{Well}}(a) \,\text{d}u \end{aligned}$$

Bibliography

[1] B Carstensen. Age-Period-Cohort models for the Lexis diagram. *Statistics in Medicine*, 26(15):3018–3045, July 2007.

[2] B. Carstensen, J.K. Kristensen, P. Ottosen, and K. Borch-Johnsen. The Danish National Diabetes Register: Trends in incidence, prevalence and mortality. *Diabetologia*, 51:2187–2196, 2008.

[3] S. A. Grover, M. Kaouache, P. Rempel, L. Joseph, M. Dawes, D. C. Lau, and I. Lowensteyn. Years of life lost and healthy life-years lost from diabetes and cardiovascular disease in overweight and obese people: a modelling study. *Lancet Diabetes Endocrinol*, 3(2):114–122, Feb 2015.

[4] L. Huo, J. L. Harding, A. Peeters, J. E. Shaw, and D. J. Magliano. Life expectancy of type 1 diabetic patients during 1997-2010: a national Australian registry-based cohort study. *Diabetologia*, 59(6):1177–1185, Jun 2016.

[5] L. Huo, J. E. Shaw, E. Wong, J. L. Harding, A. Peeters, and D. J. Magliano. Burden of diabetes in Australia: life expectancy and disability-free life expectancy in adults with diabetes. *Diabetologia*, 59(7):1437–1445, Jul 2016.

[6] M. Y. Leung, L. M. Pollack, G. A. Colditz, and S. H. Chang. Life years lost and lifetime health care expenditures associated with diabetes in the U.S., National Health Interview Survey, 1997-2000. *Diabetes Care*, 38(3):460–468, Mar 2015.

[7] S. J. Livingstone, D. Levin, H. C. Looker, R. S. Lindsay, S. H. Wild, N. Joss, G. Leese, P. Leslie, R. J. McCrimmon, W. Metcalfe, J. A. McKnight, A. D. Morris, D. W. Pearson, J. R. Petrie, S. Philip, N. A. Sattar, J. P. Traynor, and H. M. Colhoun. Estimated life expectancy in a Scottish cohort with type 1 diabetes, 2008-2010. *JAMA*, 313(1):37–44, Jan 2015.

[8] Z. Wang and M. Liu. Life years lost associated with diabetes: An individually matched cohort study using the U.S. National Health Interview Survey data. *Diabetes Res. Clin. Pract.*, 118:69–76, Aug 2016.

[9] J. C. Wright and M. C. Weinstein. Gains in life expectancy from medical interventions–standardizing data on outcomes. *N. Engl. J. Med.*, 339(6):380–386, Aug 1998.