Estimates of prediabetes and previously unknown type 2 diabetes in Denmark — the end of an epidemic or a diagnostic artefact?

Electronic Supplementary Material

SDCC
July 2018

Compiled Friday 27th July, 2018, 09:57
from: `/home/bendix/sdc/proj/DF/r/ESM.tex`

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

Contents

1

Background and theory

1.1 Introduction

If there is a *non*-differential participation by disease status in a survey, the prevalence of any condition in the source population will of course be correctly estimated by the prevalence in the survey sample. In principle independently of the size of the participation probability.

If there is a *differential* participation probability in a survey, the prevalence of a given condition will *not* be correctly estimated by the prevalence in the survey. Unfortunately, participation in health surveys is very likely to depend on severity of disease, notably on whether a person has diabetes (DM) or not.

If we know the true population prevalence of DM (from a register, say) and also have a survey, we will then be able to say something about how the participation rate depends on DM status. Ultimately we would like to use this to inform the true prevalence of diabetes, undiagnosed diabetes and pre-diabetes in the population.

If we were only interested in the prevalence of DM, the entire exercise would be superfluous, we could then just use the register prevalences.

1.2 Only DM / no DM

First, consider a very simple scenario (numbers are taken out of thin air, only for illustration); 12% prevalence of diabetes and survey response rates of 40, resp, 60% among persons with and without DM, illustrated in figure 1.1.

1.2.1 General set-up

Taking the numerical illustration in figure 1.1 to generality, suppose we have the following *known* quantities:

- unknown quantities:
  - r_1, r_2: survey participation rates for persons with resp. without DM.
  - π_1, π_2: population prevalences of DM, resp no DM; $\pi_1 + \pi_2 = 1$
- known quantities:
  - r: overall survey response rate

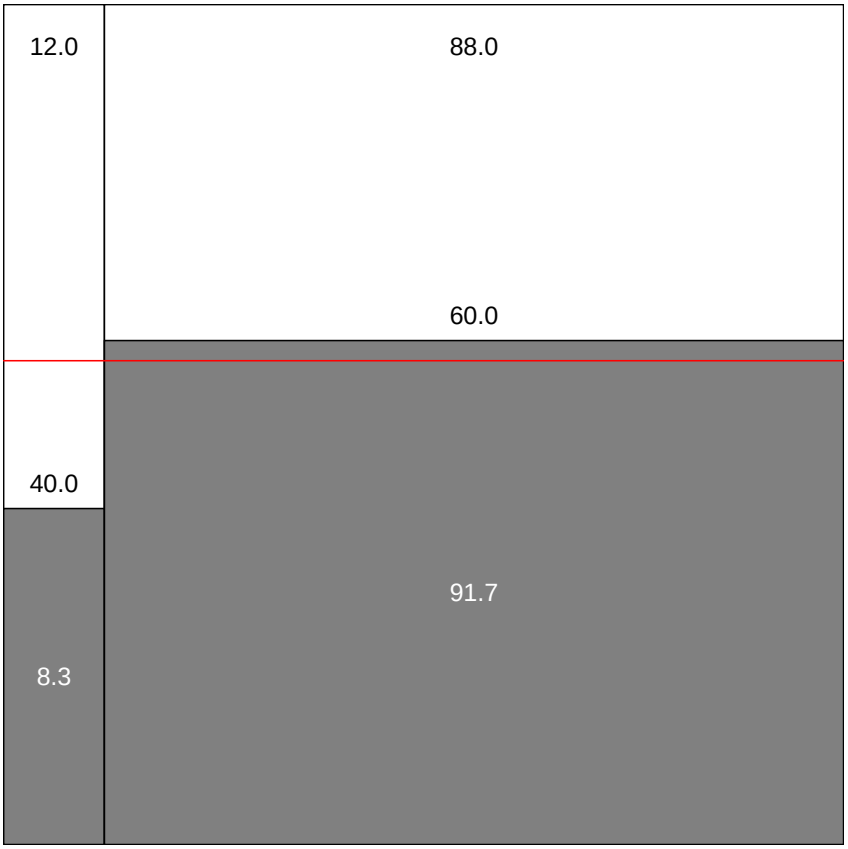

Figure 1.1: *An illustration of how differential response rates influence the result of a survey; 12% prevalence in the population, transforms to 8.3% prevalence in the surveyed part of the population. The red line is the overall survey response rate* `../graph/ESM-simple`

- p_1, p_2: survey prevalences of DM, resp no DM; $p_1 + p_2 = 1$

We have the following relations between these quantities:

$$\begin{aligned} r &= r_1\pi_1 + r_2\pi_2 \\ p_i &= r_i\pi_i/r, \qquad i = 1, 2 \end{aligned}$$

The latter is easily inverted to:

$$\pi_i = p_i r/r_i$$

So that if we know the group-specific response rates we can adjust the survey prevalences and obtain the population prevalences. However, since the group membership is only known for the surveyed, we cannot know the group-specific response rates directly.

But if we have a diabetes register, we know π_1 and $\pi_2 = 1 - \pi_1$, and so we can use this reference to compute the group specific response rates:

$$r_i = rp_i/\pi_i, \quad i = 1, 2$$

1.2.2 Age-specific prevalences

We can repeat the entire set-up above using age-specific prevalences of diabetes — this is the only realistic scenario — replacing π with $\pi(a)$ and likewise p with $p(a)$.

Now if we assume response rates r_i are constant across the age-range we still have the overall response rate r — but now dependent on age:

$$r(a) = r_1\pi_1(a) + r_2\pi_2(a)$$

But since $\pi_1(a) = 1 - \pi_2(a)$ there is no way we can maintain that the age-specific response rates are constant, so we must necessarily have:

$$r(a) = r_1(a)\pi_1(a) + r_2(a)\pi_2(a)$$

introducing a lot of extra variability, but still only with the external knowledge of the overall response rate as

$$r = \int_{20}^{85} r(a)\,\mathrm{d}a$$

where the age-range is arbitrarily taken as 20–85. Formally the expression should be:

$$r = \int_{20}^{85} r(a)f(a)\,\mathrm{d}a$$

where $f(a)$ is the density of the age-distribution in the survey sample.

1.3 Several categories of DM

Now, for illustration, suppose we have prevalence of DM (in some age class) 11%, of undetected DM 13%, of pre-diabetes 17% and consequently of no DM 61%, and further assume that response rates in a survey is 30, 40, 50 and 55% for the four categories. Asking 1000 people to participate would then produce: where we see that the groups with response rates *smaller* than the overall rate has a smaller survey prevalence than population prevalence, and vice versa.

We can illustrate this in figure 1.2: However, what is known from the survey is only the distribution of the responders in the 4 groups, indicated by the numbers in white, and the overall response rate — indicated by the red line in figure 1.2. What we are after is the population distribution in the three categories — the numbers in the top of the figure.

Note that the survey prevalence is smaller than the population prevalence for groups where the response rate is smaller than the overall response rate and vice versa.

The overall response rate is:

$$r = r_1\pi_1 + r_2\pi_2 + r_3\pi_3 + r_4\pi_4$$

and from this we have the following relationships between response rates (r_i), survey prevalences (p_i — known from the survey) and population prevalences (π_i)

$$p_i = r_i\pi_i/r \quad \Leftrightarrow \quad \pi_i = rp_i/r_i \quad \Leftrightarrow \quad r_i = rp_i/\pi_i$$

From the survey we have the overall response rates and from the DM register we have an estimate of π_1 so also an estimate of r_1. Therefore we could use information on r which we

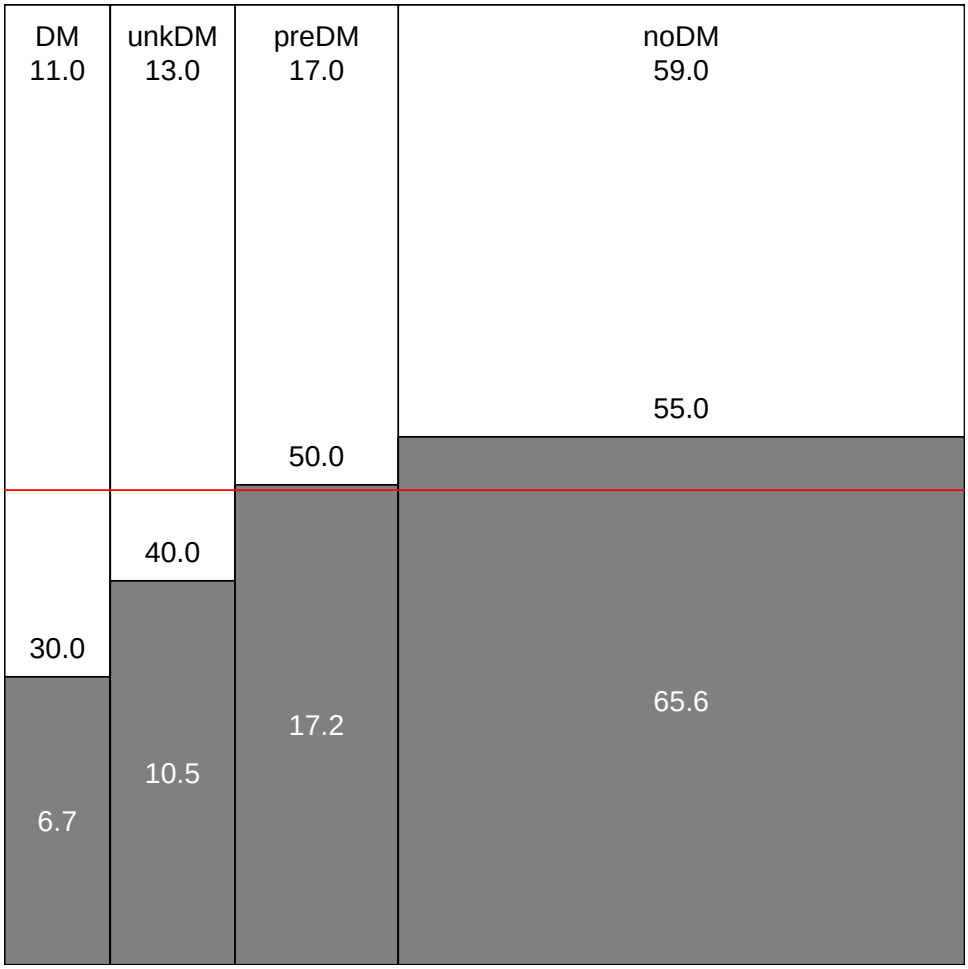

Figure 1.2: *Illustration of the result of a survey with differing response rates in different DM classes; numbers are taken out of thin air purely for illustration purposes. At the top is the population prevalences, corresponding to the* width *of the boxes. In each category is indicated with the gray area the class-specific response rates; the white numbers are the observed prevalences in the survey sample. The overall response rate is shown as the horizontal red line; the gray area above the red line is the same as the white area below the red line.../graph/ESM-groups*

also have, to guess at r_2, r_3 and r_4, which by the middle equation would give us estimates of the population prevalences, π_i. However, this will largely be guesswork, many choices of r_2, r_3 and r_4 can give estimates of the population prevalences (obeying $\sum_i \pi_i = 1$), that would make the overall response rate fit to the observed.

1.3.1 Age-dependence in the 4 classes

Clearly the prevalences depend on age, but the prevalences in the 4 classes always sum to 1, so for any age a:

$$1 = \pi_1(a) + \pi_3(a) + \pi_3(a) + \pi_4(a) \quad \text{(only } \pi_1 \text{ known)}$$
$$1 = p_1(a) + p_3(a) + p_3(a) + p_4(a) \quad \text{(all } p\text{s known)}$$

From the equations above we have the exact same relationships as above, but now by age:

$$p_i(a) = r_i(a)\pi_i(a)/r(a) \quad \Leftrightarrow \quad \pi_i(a) = r(a)p_i(a)/r_i(a) \quad \Leftrightarrow \quad r_i(a) = r(a)p_i(a)/\pi_i(a)$$

where $r(a) = \sum_j r_j(a)\pi_j(a)$ is the overall response rate at age a.

If we make the bold assumption that the *ratio* of the class-specific response rates to the overall response rate $r(a)$ is constant over age — we assume they vary the same way with age, the probability for a person aged a of being in the survey and classified in class i is:

$$p_i(a) = \pi_i(a)\frac{r_i(a)}{r(a)} = \pi_i(a)k_i \quad \Leftrightarrow \quad k_i = \frac{r_i(a)}{r(a)}, \forall a$$

where k_i the ratio of the ith class-specific response rate to the overall assumed overall response rate, assumed to be constant across ages.

If we have expressions for $p_i(a)$ (derived from the survey data) and a value of k_1 we have three free parameters to manipulate: k_2, k_3 and k_4. For any given combination of these we can compute the population prevalences, using $\pi_i(a) = p_i(a)/k_i$. Referring to figure 1.2, we would in practice expect that $k_1 < 1$ and that $k_1 < k_2 \leq k_3 \leq k_4$. In the practical calculations we shall impose the latter assumption if we observe the former (as we have data to do). Actually, we will use the (fairly arbitrary) restriction that $k_i = k_1 + w_i\kappa$ for a fixed set of numbers w_i with $w_1 = 0$ and $w_2 = 1$, in order to reduce the problem to a one-parameter problem that can be handled in practice. In practice we will use $w = (0, 1, 1.6, 2)$.

We can use the survey data to model $p_i(a)$ as a function of age; this function will be proportional to the population prevalence as function of age by the assumption of constant (age-independent) ratios $k_i = r_i(a)/r(a)$. We have that the overall observed response rate at age a, for a given set of values of k_i is:

$$r(a) = k_1 r(a)\pi_1(a) + k_2 r(a)\pi_2(a) + k_3 r(a)\pi_3(a) + k_4 r(a)\pi_4(a) = r(a)\sum_{i=1}^{4} k_i\pi_i(a)$$

Thus, for the chosen k_is to be credible, $r(a)$ integrated with respect to the age distribution in the survey should be equal to the overall observed response rate, r. But we do not know the true $r(a)$, so another needed assumption would be that $r(a) = r$, independent of age, but we actually only need to assume that:

$$\int_a r(a)\sum_{i=1}^{4} k_i\pi_i(a) = r\int_a \sum_{i=1}^{4} k_i\pi_i(a)$$

that is

$$\int_a \sum_{i=1}^{4} k_i\pi_i(a) = 1$$

where the integration is over the empirical age-distribution *in the survey.*

If we model the survey prevalences (p_i) correctly, one possible solution is of course $k_1 = k_2 = k_3 = k_4 = 1$ — the trivial solution, but from the DM register we have a clue as to what k_1 is (certainly not 1!), from the relation

$$p_1(a) = k_1\pi_1(a) \quad \Leftrightarrow \quad k_1 = \frac{p_1(a)}{\pi_1(a)}$$

where we know $p_1(a)$ from the survey data and $\pi_1(a)$ from the register data.

1.4 Algorithm for deriving population prevalences

The considerations above lead to the following algorithm for correcting the observed age-specific prevalences in the surveys, using the overall response rate and the age-specific prevalences of DM from the register.

The steps to derive the corrected population prevalences are as follows (using the numbers 1,2,3,4 for the states DM, unkDM, preDM and noDM respectively):

1. Estimate k_1 from the survey data by:

   (a) Fit a log-link binomial `glm` for the *register* prevalence $\pi_1(a,t)$ of known DM by age and calendar time (and possibly birth cohort...) — producing a smooth function of age and calendar time.

   (b) Use this to predict the log-age-specific (register based) prevalence of known DM, $\pi_1(a,t)$ at the age, date (a,t) points of the *survey* (that is on the linear predictor scale).

   (c) Fit a log-link binomial `glm` to the indicator of DM in the *survey* (the survey prevalence of DM, $p_1(a,t)$), with only intercept and with the predicted log-prevalences of DM from the register as offset.

   (d) The only parameter in this model (the intercept, μ) will be $\log(k_1)$, because the model for the survey probability of DM is:

   $$\log\big(p_1(a,t)\big) = \mu + \log\big(\hat{\pi}_1(a,t)\big) \quad \Leftrightarrow \quad p_1(a,t) = \mathrm{e}^{\mu}\hat{\pi}_1(a,t) = k_1\hat{\pi}_1(a,t) \quad \Leftrightarrow \quad k_1 = \mathrm{e}^{\mu}$$

   Since some of the surveys are conducted over a period of some years, we also incorporated the calendar time effect in the predicted prevalence of known diabetes ($\hat{\pi}_1$) in the equation above.

   This way the crucial assumption that $k_1 = r_1(a)/r(a)$ is independent of a is implemented in the modeling via non-inclusion of age in the model for survey prevalences and by using the offset of the linear predictor from the model for the register based prevalence of DM.

2. Fit models for the survey prevalences $p_i(a), i = 1, \ldots 4$ of known-DM, unkn-DM, pre-DM, as smooth functions of age. We have fitted these by fitting the prevalence of known-DM, of known-DM + unkn-DM and of known-DM + unkn-DM + pre-DM respectively, and taking the difference between these as the needed.

3. Choose k_2, k_3, k_4. These are the parameters that we can manipulate in order to allow for differential response rates while keeping the overall response rate as observed (which it should be).

   In order to arrive at a one-dimensional optimization problem, we assume that k_1 is the smallest and that $k_i = k_1 + w_i \times \kappa$ for some κ and a *fixed* set of (arbitraily chosen) values for w_i. The optimization problem is then to determine κ.

4. For the now defined k_is, we compute $\tilde{\pi}_2(a)$, $\tilde{\pi}_3(a)$ and $\tilde{\pi}_4(a)$ from the ks and the fitted ps: $\tilde{\pi}_i(a) = p_i(a)/k_i$. Note that there is no guarantee here that $\sum_j \tilde{\pi}_j(a) = 1, \forall a$ from this procedure, as it should be.

5. Hence we adjust the π_i to sum 1 at any age, keeping the register-based $\pi_1(a)$:

$$\pi_i(a) = \left(\tilde{\pi}_i(a) \Big/ \sum_{j=2}^{4} \tilde{\pi}_j(a) \right) \times \big(1 - \pi_1(a)\big), \; i = 2, 3, 4$$

   Thus we have a set of age-specific partitions of the population in the four groups, based on the k_1 and the arbitrarily chosen κ.

6. Check that the observed total response rate fits with the predicted by checking that $\int_a \sum_i k_i \pi_i(a) \, \mathrm{d}a = 1$

7. Adjust the k_is; that is κ, to obtain this if it is not the case.

   The latter is achieved by putting the entire algorithm outlined (steps 2–6) into a function with κ as argument and $\int_a \sum_i k_i \pi_i(a) \, \mathrm{d}a - 1$ as result and then using `uniroot` to find the κ that returns 0.

Note that this last point constitutes the iteration / optimization. Basically, the (relationship between) w_i parameters is taken out of thin air, and only κ is adjusted so that the overall response rate fits with the constructed πs.

2

Data acquisition

2.1 Survey data sources

We load the dataframe with the survey data from the five different surveys:

```
> load( file="../data/tot.Rda" )
> # summary( tot )
> with( tot, ftable( addmargins( table( sex, st, dmst, useNA="ifany" ) ),
+                    row.vars = 1:2 ) )
           dmst known-DM unkn-DM pre-DM  Well    NA   Sum
sex st
M   I-99             72      57    225  2946     2  3302
    DANHES          209      47    361  6297   446  7360
    H-06             77      14     61  1382    19  1553
    H-08              0       2     12   330     2   346
    GESUS           521     199    852  7960     0  9532
    Sum             879     319   1511 18915   469 22093
F   I-99             67      25    127  3257     6  3482
    DANHES          206      50    451  9428   570 10705
    H-06             65       9     81  1743    20  1918
    H-08              5       1     24   417     2   449
    GESUS           409     136    938  9926     0 11409
    Sum             752     221   1621 24771   598 27963
Sum I-99            139      82    352  6203     8  6784
    DANHES          415      97    812 15725  1016 18065
    H-06            142      23    142  3125    39  3471
    H-08              5       3     36   747     4   795
    GESUS           930     335   1790 17886     0 20941
    Sum            1631     540   3132 43686  1067 50056
> with( tot, tapply( hba, list(dm,dmst), min, na.rm=T ) )
  known-DM  unkn-DM   pre-DM     Well
Y 2.543449       NA       NA       NA
N       NA 6.450485 5.992987 2.497699
> with( tot, tapply( hba, list(dm,dmst), max, na.rm=T ) )
  known-DM unkn-DM pre-DM     Well
Y     15.3      NA     NA       NA
N       NA    13.9    6.4 5.901487
```

We want a brief overview of when and in what ages persons were surveyed in the different studies;

It is clear from figure 2.1 that the time trend in the occurrence of the various degrees of DM based on all 4 studies is going to be based on the difference between the Inter-99 study and at most the three other studies (DANHES, Helbred-2008 and GESUS).

Finally we save the survey dataset along with a vector of response rates, for use in further analyses:

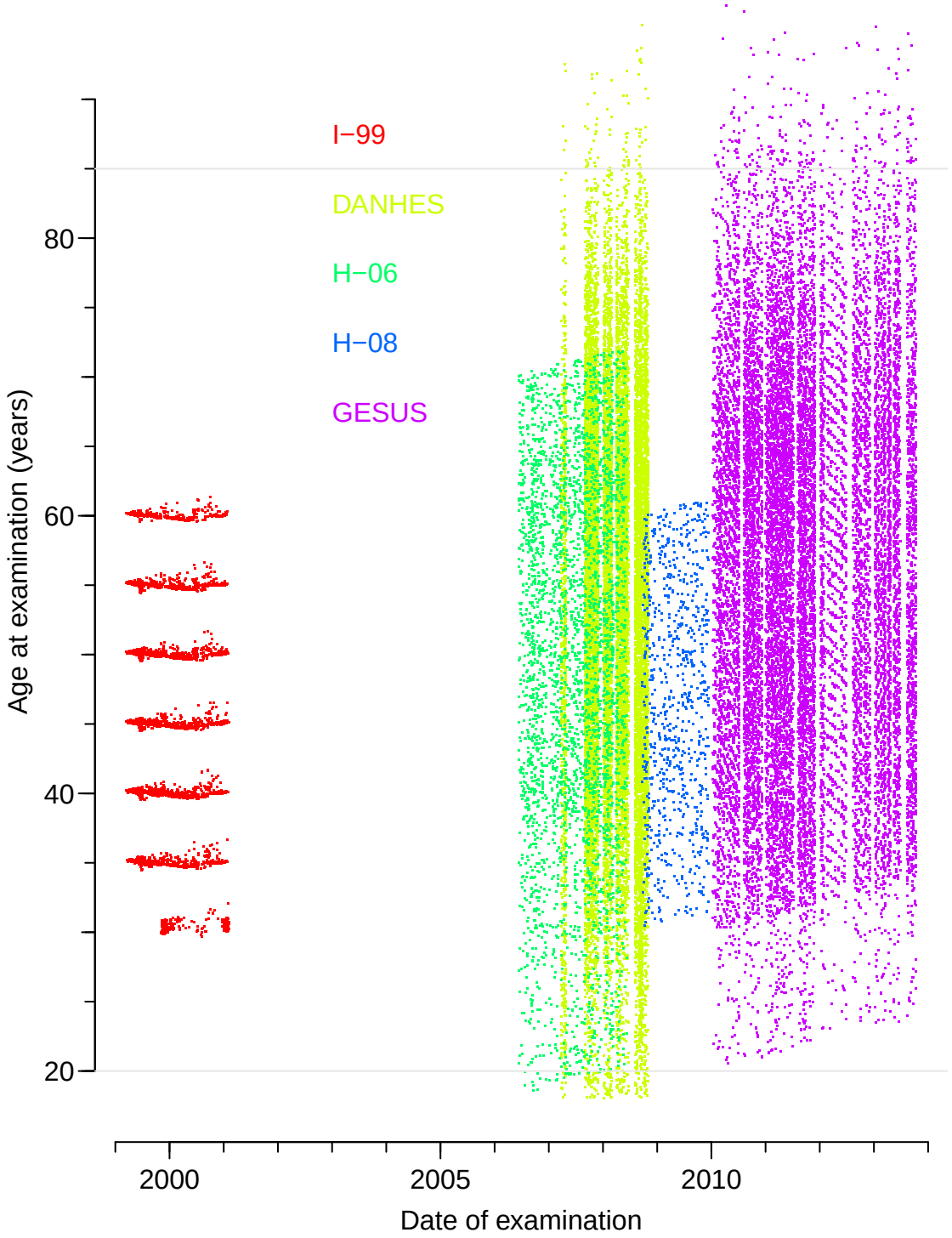

Figure 2.1: *Date and age at examination for the studies in this analysis.* `../graph/ESM-exdat`

2.2 Register prevalence of T2D

We finally read the data from the reconstructed diabetes register: We now have prevalent number of T2D cases (`X`) and number of persons without diabetes (`N`) alive at each combinnation of sex (`sex`), age (`A`) and date (`P`); the latter two in 1 year age classes.

Ignoring the fact that persons appear several times in the dataset, we model the prevalence of diabetes by a log-link binomial model with spline effects of age period and cohort, in order to be able to make resinably accurate predictions of prevalence (the probability that a given person has diabetes) for any combination of age and date:

```
> ( a.kn <- seq(20,85,,9) )
[1] 20.000 28.125 36.250 44.375 52.500 60.625 68.750 76.875 85.000
> ( p.kn <- seq(1996,2015,,5) )
[1] 1996.00 2000.75 2005.50 2010.25 2015.00
> ( c.kn <- seq(1900,2010,,9) )
[1] 1900.00 1913.75 1927.50 1941.25 1955.00 1968.75 1982.50 1996.25 2010.00
> mm <- glm( cbind(X,N) ~ Ns(  A,kn=a.kn) +
+                         Ns(P  ,kn=p.kn) +
+                         Ns(P-A,kn=c.kn),
+                 family = binomial("log"),
+                   data = subset(prev,sex=="M") )
> mw <- update( mm, data = subset(prev,sex=="F") )
```

Now we have the two model objects `mm` (men) and `mw` (women) with prevalences that we can use for prediction. We briefly illustrate the predicted prevalences at selected dates: Finally we save the model objects for further use:

2.2.1 The limited age-range

All calculations will be for persons in the age-range 20–85, because this is the range reliably covered by the surveys. However we will make bold extrapolations for the age-range over 85, and under 20, so it is useful to see the populations size in these age-brackets:

```
> data( N.dk )
> # Age-groups
> N.dk$Ag <- cut( N.dk$A, breaks=c(-Inf,20,85,Inf), right=FALSE )
> # Table by sex, period, age group with margin
> tt <- addmargins( xtabs( N~sex+P+Ag, data=subset(N.dk,P>2006) ) )
> # utility for nice printing:
> fftable <-
+ function( tt, w, d, ... )
+ ftable( formatC( as.table(tt), # keeps the original dimensions of ftable objects
+                  format="f", width=w, digits=d, big.mark="," ),
+          ... )
> # print the tables of totals and percentages properly:
> fftable( tt, w=12, d=0 )
         Ag    [-Inf,20)      [20,85)    [85, Inf)          Sum
sex P
1   2007         685,423    1,979,885       31,354    2,696,662
    2008         688,870    1,991,730       32,066    2,712,666
    2009         692,283    2,007,051       32,686    2,732,020
    2010         693,000    2,016,744       33,542    2,743,286
```

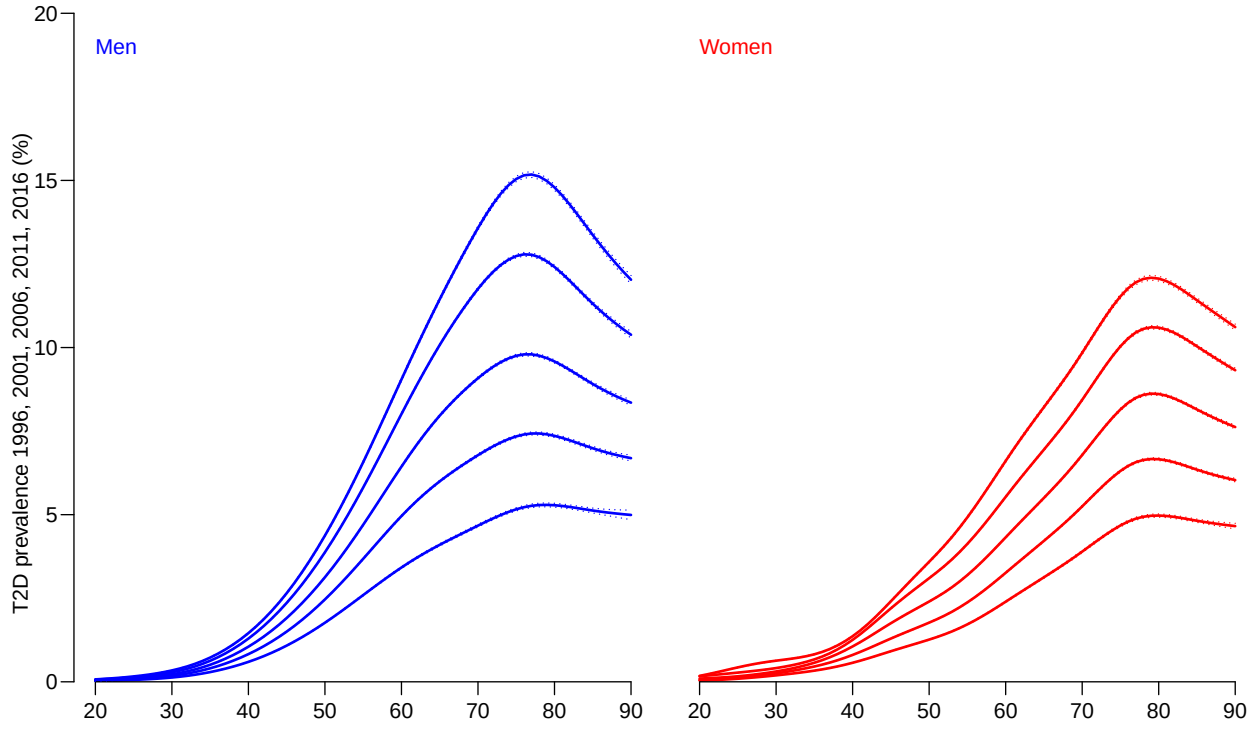

Figure 2.2: *Predicted prevalences of T2D from the reconstructed diabetes register (using an age-period-cohort model for the prevalences classified in 1-year age-classes) for each 1st January 1996, 2001, 2006, 2011, 2016, with 95% confidence limits.* `../graph/ESM-prev`

```
     2011         692,534    2,029,675       34,373    2,756,582
     2012         689,402    2,041,890       35,484    2,766,776
     2013         684,537    2,057,945       36,370    2,778,852
     Sum        4,826,049   14,124,920      235,875   19,186,844
2    2007         651,551    2,024,370       74,501    2,750,422
     2008         655,310    2,033,037       74,778    2,763,125
     2009         658,730    2,044,931       75,770    2,779,431
     2010         659,246    2,055,785       76,421    2,791,452
     2011         659,019    2,068,084       76,943    2,804,046
     2012         656,181    2,080,002       77,557    2,813,740
     2013         651,283    2,094,350       78,143    2,823,776
     Sum        4,591,320   14,400,559      534,113   19,525,992
Sum  2007       1,336,974    4,004,255      105,855    5,447,084
     2008       1,344,180    4,024,767      106,844    5,475,791
     2009       1,351,013    4,051,982      108,456    5,511,451
     2010       1,352,246    4,072,529      109,963    5,534,738
     2011       1,351,553    4,097,759      111,316    5,560,628
     2012       1,345,583    4,121,892      113,041    5,580,516
     2013       1,335,820    4,152,295      114,513    5,602,628
     Sum        9,417,369   28,525,479      769,988   38,712,836
> fftable( tt/tt[,,rep(4,4)]*100, w=7, d=1 )
          Ag [-Inf,20) [20,85) [85, Inf)     Sum
sex P
1    2007         25.4    73.4       1.2   100.0
     2008         25.4    73.4       1.2   100.0
     2009         25.3    73.5       1.2   100.0
```

```
    2010         25.3    73.5       1.2   100.0
    2011         25.1    73.6       1.2   100.0
    2012         24.9    73.8       1.3   100.0
    2013         24.6    74.1       1.3   100.0
    Sum          25.2    73.6       1.2   100.0
2   2007         23.7    73.6       2.7   100.0
    2008         23.7    73.6       2.7   100.0
    2009         23.7    73.6       2.7   100.0
    2010         23.6    73.6       2.7   100.0
    2011         23.5    73.8       2.7   100.0
    2012         23.3    73.9       2.8   100.0
    2013         23.1    74.2       2.8   100.0
    Sum          23.5    73.8       2.7   100.0
Sum 2007         24.5    73.5       1.9   100.0
    2008         24.5    73.5       2.0   100.0
    2009         24.5    73.5       2.0   100.0
    2010         24.4    73.6       2.0   100.0
    2011         24.3    73.7       2.0   100.0
    2012         24.1    73.9       2.0   100.0
    2013         23.8    74.1       2.0   100.0
    Sum          24.3    73.7       2.0   100.0
```

```
> # Restrict to people over 60:
> tt <- addmargins( xtabs( N~sex+P+Ag, data=subset(N.dk,P>2006 & A>60) ), 3 )
> fftable( tt/tt[,,rep(4,4)]*100, w=7, d=1 )
```

```
          Ag [-Inf,20) [20,85) [85, Inf)     Sum
sex P
1   2007           0.0    93.7       6.3   100.0
    2008           0.0    93.8       6.2   100.0
    2009           0.0    93.9       6.1   100.0
    2010           0.0    93.9       6.1   100.0
    2011           0.0    93.9       6.1   100.0
    2012           0.0    93.8       6.2   100.0
    2013           0.0    93.8       6.2   100.0
2   2007           0.0    87.9      12.1   100.0
    2008           0.0    88.1      11.9   100.0
    2009           0.0    88.2      11.8   100.0
    2010           0.0    88.3      11.7   100.0
    2011           0.0    88.4      11.6   100.0
    2012           0.0    88.5      11.5   100.0
    2013           0.0    88.6      11.4   100.0
```

Thus we see that men over 85 is about 1.2% of the entire population, and about 6.2% of the population over 60, whereas the corresponding figures for women are 2.7% and 11.5% — substantially more, reflecting the longer lifespan of women.

3

Analysis correcting for response

In this chapter, we implement the algorithm described in section 1.4.

3.1 Data and T2D register prevalence

First we load the data and the relevant analysis package

```
> library( Epi )
> clear()
> load( file="../data/tot.Rda" )
```

The outcome variable of interest is the persons' diabetes status — a 4-level ordered categorical factor.

```
> with( tot, addmargins(table(st,dmst)) )
        dmst
st       known-DM unkn-DM pre-DM  Well   Sum
  I-99        139      82    352  6203  6776
  DANHES      415      97    812 15725 17049
  H-06        142      23    142  3125  3432
  H-08          5       3     36   747   791
  GESUS       930     335   1790 17886 20941
  Sum        1631     540   3132 43686 48989
> with( tot, table(doe-dob>85,st) )
       st
        I-99 DANHES  H-06  H-08 GESUS
  FALSE 6784  17992  3471   795 20710
  TRUE     0     73     0     0   231
```

The practical modeling is done by log-link binomial regression of outcomes defined as cumulative successive levels; that is 1st level (known T2D), at most 2nd level (known + unknown DM) and at most 3rd level (known + unknown + pre DM).

We will not include the Inter99 study, because it is too old:

```
> tot <- subset( tot, st !="I-99" )
> tot$st <- factor( tot$st )
> cbind( with( tot, table( st, dmst, useNA="ifany" ) ), resr )
       known-DM unkn-DM pre-DM  Well <NA>  resr
DANHES      415      97    812 15725 1016 0.140
H-06        142      23    142  3125   39 0.447
H-08          5       3     36   747    4 0.440
GESUS       930     335   1790 17886    0 0.427
```

```
> with( tot, pctab( table( st, dmst, useNA="ifany" ) ) )
        dmst
st       known-DM unkn-DM  pre-DM    Well    <NA>     All       N
  DANHES      2.3     0.5     4.5    87.0     5.6   100.0 18065.0
  H-06        4.1     0.7     4.1    90.0     1.1   100.0  3471.0
  H-08        0.6     0.4     4.5    94.0     0.5   100.0   795.0
  GESUS       4.4     1.6     8.5    85.4     0.0   100.0 20941.0
```

We will also need the fitted population prevalences that was modeled previously; in the model objects `mm` and `mw` for men and women, respectively:

```
> load( file="../data/prmod.Rda" )
> lls()
  name mode    class      dim           size(Kb)
1 a.kn numeric numeric    9                 0.2
2 c.kn numeric numeric    9                 0.2
3 mm   list    glm lm     30            1,945.8
4 mw   list    glm lm     30            1,945.8
5 p.kn numeric numeric    5                 0.1
6 resr numeric numeric    4                 0.4
7 tot  list    data.frame 43272 7       4,060.1
```

3.2 Algorithm

First we define the function `mfit` to fit the binomial models for the survey prevalences. The purpose is to derive k_1 and predicted values of the age-specific prevalences at 1) the survey ages and 2) a set of equidistant prediction ages; the latter for reporting purposes. Also we want the median date of survey.

The input to the function is:

- `dfr` — the survey dataframe; it is required that variables `A` (age at survey) and `P` (date of survey) are in the data frame.
- `popprv` — a fitted binomial model for the register prevalence of DM, as a function of `A` and `P`. This will be either `mm` or `mw`.
- `lo`, `hi`, `il` — specification of the age points for prediction of prevalences, they will be midponts of intervals of length `il` between ages `lo` and `hi`.
- `a.kn` — placement of the knots on the age-scale for the models of prevalences in the survey.

The function returns a list with components:

- `k1` — the ratio of the DM response rate to the overall response rate
- `prvp` — the log-fitted survey proportions at the ages (and dates) in the survey for the four groups in an $n_{\text{survey}} \times 4$ matrix (n_{survey} is the number of persons in the survey).
- `prvs` — the log-fitted survey proportions by equidistant ages (the chosen age-prediction points) for the four groups in a $n_a \times 4$ matrix (n_a is the number of prediction points for age).

- `srvd` — the median date of survey. This is to be used for prediction of the *number* of persons in the different groups; we will use the population size interpolated to the median date of survey.

```
> mfit <-
+ function( dfr, popprv,
+           lo=20, hi=85, il=1,
+           a.kn=3:8*10 )
+ {
+ nd <- dfr[,c("A","P")]
+ prls <- predict( popprv, newdata=nd, type="link" )
+ # Find k1
+ mod.k1 <- glm( ( dmst==levels(dfr$dmst)[1] ) ~ 1,
+                 offset = prls,
+                 family = binomial( link="log" ),
+                 data = dfr )
+ k1 <- exp( coef(mod.k1) )
+ # Fit the age-specific prevalences of at least unkn-DM, resp. pr-DM
+ mod1 <- glm( (dmst %in% levels(dfr$dmst)[1  ]) ~ Ns(A,knots=a.kn),
+              family = binomial( link="log" ),
+              data = dfr )
+ mod2 <- glm( (dmst %in% levels(dfr$dmst)[1:2]) ~ Ns(A,knots=a.kn),
+              family = binomial( link="log" ),
+              data = dfr )
+ mod3 <- glm( (dmst %in% levels(dfr$dmst)[1:3]) ~ Ns(A,knots=a.kn),
+              family = binomial( link="log" ),
+              data = dfr )
+ # Age midpoints in intervals between lo and hi
+ pr.a <- seq(lo,hi,il)[-1] - il/2
+ pr.p <- median(dfr$doe)
+ pdat <- data.frame(A=pr.a,P=pr.p)
+ prlp <- predict( popprv, newdata=pdat, type="link" )
+ # here are the p_i functions at the prediction points
+ prvp <- exp( cbind( predict( mod1, newdata=pdat, type="link" ),
+                     predict( mod2, newdata=pdat, type="link" ),
+                     predict( mod3, newdata=pdat, type="link" ) ) )
+ prvp <- cbind( prvp[,1], prvp[,2]-prvp[,1], prvp[,3]-prvp[,2], 1-prvp[,3] )
+ # here are the p_i functions at the survey points
+ prvs <- exp( cbind( predict( mod1, newdata=nd, type="link" ),
+                     predict( mod2, newdata=nd, type="link" ),
+                     predict( mod3, newdata=nd, type="link" ) ) )
+ prvs <- cbind( prvs[,1], prvs[,2]-prvs[,1], prvs[,3]-prvs[,2], 1-prvs[,3] )
+ colnames( prvp ) <- colnames( prvs ) <- levels( dfr$dmst )
+ rownames( prvp ) <- pdat$A
+ rownames( prvs ) <-   nd$A
+ list( k1=k1, prvp=prvp, prvs=prvs, srvd=pr.p,
+       prvDM.pred=exp(prlp),
+       prvDM.surv=exp(prls) )
+ }
```

Next we define a function that takes the output from `mfit`, makes a bold guess at the group specific response rates and returns the deviation between the overall predicted response rate and the actually observed (which should be 0), as well as the corrected age-specific prevalences in the three (well, four) groups.

The function `cprv` takes the following input:

- `inc` — the increment from k_1 to k_2
- `shp` — the shape of the increments over the groups; multiplied with `inc` to give the ks. It is anticipated that the first two elements of `shp` are 0, 1 (this is not checked, though) `inc` and `shp` are merely used to define $k_i = k_1 + \texttt{inc} * \texttt{shp}[i], \texttt{i} = 2, 3, 4$
- `dfr` — the survey data set
- `totres` — the overall response rate in the survey
- `mfmod` — a list as returned from the function `mfit`.

The function `cprv` constructs group specific response rates and use these to construct and return the following objects in a list:

- `respr` — a 4-vector of response rates for each of the four diagnostic groups.
- `prvp` — the corrected age-specific prevalences for the four groups at the prediction ages from `mfit`.
- `prvs` — the original survey probabilities at the prediction ages from `mfit`.
- `dev` — the difference between the observed total response rate and the response rate computed from the corrected population prevalences. The purpose of this is to be able to iterate over values of `inc` to find a value which together with the (rather arbitraily chosen) `shp` gives reasonably corrected category-specific response rates.

```
> cprv <-
+ function( inc = 0.2,        # how much does class-specific response rates change
+           shp=c(0,1,1.6,2), # and in what shape?
+           dfr, totres,      # survey data set and total response rate
+           mfmod )
+ {
+ kk <- mfmod$k1 + inc * shp
+ names( kk ) <- levels( dfr$dmst )
+ # once the group-specific ks are chosen we can compute the group
+ # specific response rates from the overall response rate rt
+ rr <- kk * totres
+ # devise the pi_i functions at both prediction and survey points
+ pi.s <- sweep( mfmod$prvs, 2, kk, "/" )
+ pi.p <- sweep( mfmod$prvp, 2, kk, "/" )
+ # however, we know the DM prevalences in population so those are kept
+ pi.s[,1] <- mfmod$prvDM.surv
+ pi.p[,1] <- mfmod$prvDM.pred
+ # adjust pi for the remaining 3 categories so age-specific sums are 1
+ colnrm <- function( M ) cbind( M[, 1],
+                          sweep( M[,-1], 1, apply(M[,-1],1,sum)/(1-M[,1]), "/" ) )
+ fits <- colnrm( pi.s )
+ fitp <- colnrm( pi.p )
+ # how does that leave the overall empirical response rates
+ prr <- mean( fits %*% rr )
+ # return the results
+ list( respr = rr,
+        prvp = fitp, #
```

```
+        prvs = mfmod$prvp, # survey probabilities evaualted at predction ages.
+        srvd = mfmod$srvd,
+         dev = prr-totres )
+ }
```

Finally we devise a wrapper, `findprv`, that puts it all to uniroot and returns the relevant results: the original and corrected prevalences of the four classes and the derived class response rates. It assumes that the overall response rates, `resr` as well as the models for register rates for men and women, `mm`, resp. `mw` are in the global environment.

```
> findprv <-
+ function( is,  # sex
+           iu,  # study
+          shp = c(0,1,1,6,2) ) # shape of non-repsons
+ {
+ dfr <- transform( subset( tot, sex==is & st==iu ),
+                   A = doe - dob,
+                   P = doe )
+ mfmod <- mfit( dfr,
+             popprv = if( is=="M") mm else mw,
+             lo=20, hi=85, il=1 )
+ uf <- uniroot( function(x) cprv( inc = x,
+                                  shp = shp,
+                                  dfr = dfr,
+                                totres = resr[iu],
+                                 mfmod = mfmod )$dev, 0:1 )
+ res <- cprv( inc=uf$root, shp=shp, dfr=dfr, totres=resr[iu], mfmod=mfmod )
+ cat( "dev(", iu, ",", is, ")=", res$dev, "\n" )
+ list( prvp = res$prvp,
+       prvs = res$prvs,
+       srvd = res$srvd,
+      respr = res$respr )
+ }
```

These functions now enables the analysis for the 4 combinations of sex and surveys of interest (DANHES and GESUS).

3.3 Analyses of surveys

We collect the predicted prevalences and revised response rates from all studies in designed arrays:

```
> lo <- 20
> hi <- 85
> il <- 1
> prarr <- NArray( list( respl = c("2.0-3.0","1.6-2.0","1.3-1.5","1.1-1.2"),
+                        study = levels(tot$st)[c(1,4)],
+                          sex = levels(tot$sex),
+                          age = seq(lo,hi,il)[-1]-il/2,
+                         type = c("Survey","Pop"),
+                          grp = levels(tot$dmst) ) )
> rrarr <- NArray( dimnames(prarr)[c(1:3,6)] )
> dsarr <- NArray( dimnames(prarr)[1] )
> str( prarr )
```

```
 logi [1:4, 1:2, 1:2, 1:65, 1:2, 1:4] NA NA NA NA NA NA ...
 - attr(*, "dimnames")=List of 6
  ..$ respl: chr [1:4] "2.0-3.0" "1.6-2.0" "1.3-1.5" "1.1-1.2"
  ..$ study: chr [1:2] "DANHES" "GESUS"
  ..$ sex  : chr [1:2] "M" "F"
  ..$ age  : chr [1:65] "20.5" "21.5" "22.5" "23.5" ...
  ..$ type : chr [1:2] "Survey" "Pop"
  ..$ grp  : chr [1:4] "known-DM" "unkn-DM" "pre-DM" "Well"
> str( rrarr )
 logi [1:4, 1:2, 1:2, 1:4] NA NA NA NA NA NA ...
 - attr(*, "dimnames")=List of 4
  ..$ respl: chr [1:4] "2.0-3.0" "1.6-2.0" "1.3-1.5" "1.1-1.2"
  ..$ study: chr [1:2] "DANHES" "GESUS"
  ..$ sex  : chr [1:2] "M" "F"
  ..$ grp  : chr [1:4] "known-DM" "unkn-DM" "pre-DM" "Well"
> str( dsarr )
 logi [1:4(1d)] NA NA NA NA
 - attr(*, "dimnames")=List of 1
  ..$ respl: chr [1:4] "2.0-3.0" "1.6-2.0" "1.3-1.5" "1.1-1.2"
```

With the results array set up we can loop over the shapes, the surveys and the two sexes:

```
> for( ir in dimnames(prarr)[[1]] )
+ for( iu in dimnames(prarr)[[2]] )
+ for( is in dimnames(prarr)[[3]] )
+ {
+ sh <- c(0,1,as.numeric(strsplit(ir,"-")[[1]]))
+ res <- findprv( is, iu, shp=sh )
+ prarr[ir,iu,is,,"Survey",] <- res$prvs
+ prarr[ir,iu,is,,"Pop"   ,] <- res$prvp
+ rrarr[ir,iu,is,] <- res$respr
+ dsarr[iu] <- res$srvd
+ }
dev( DANHES , M )= 1.251982e-06
dev( DANHES , F )= 7.906891e-07
dev( GESUS , M )= 6.663469e-06
dev( GESUS , F )= 5.475869e-06
dev( DANHES , M )= 1.322471e-07
dev( DANHES , F )= 8.215291e-08
dev( GESUS , M )= 9.152734e-07
dev( GESUS , F )= 6.689666e-07
dev( DANHES , M )= 1.548126e-08
dev( DANHES , F )= 9.801426e-09
dev( GESUS , M )= 1.223637e-07
dev( GESUS , F )= 8.808506e-08
dev( DANHES , M )= -1.974847e-06
dev( DANHES , F )= -3.494908e-06
dev( GESUS , M )= 8.013421e-09
dev( GESUS , F )= 6.322666e-09
```

We can now extract the estimated response rates for the two surveys in question:

```
> rrarr <- rrarr[,,,c(1:4,4)] * 100
> dimnames(rrarr)[[4]][5] <- "Survey"
> rrarr[,,,5] <- resr[rep(dimnames(rrarr)[[2]],2,each=dim(prarr)[1])]*100
> round( ftable(rrarr), 1 )
```

```
                  grp known-DM unkn-DM pre-DM Well Survey
respl   study  sex
2.0-3.0 DANHES M           7.7    10.0   12.3 14.5   14.0
                F          7.8    10.0   12.2 14.4   14.0
        GESUS  M          33.9    37.2   40.5 43.9   42.7
                F         30.9    35.2   39.5 43.8   42.7
1.6-2.0 DANHES M           7.7    11.1   13.1 14.5   14.0
                F          7.8    11.0   13.0 14.3   14.0
        GESUS  M          33.9    38.8   41.7 43.7   42.7
                F         30.9    37.3   41.1 43.6   42.7
1.3-1.5 DANHES M           7.7    12.2   13.5 14.4   14.0
                F          7.8    12.1   13.4 14.3   14.0
        GESUS  M          33.9    40.3   42.3 43.6   42.7
                F         30.9    39.3   41.8 43.5   42.7
1.1-1.2 DANHES M           7.7    13.3   13.8 14.4   14.0
                F          7.8    13.2   13.7 14.3   14.0
        GESUS  M          33.9    41.9   42.7 43.5   42.7
                F         30.9    41.3   42.4 43.4   42.7
```

Re-arranging the order to compare the shape of the non-responder shapes the very small deviations in the estimated response rates:

```
> round( ftable(rrarr, row.vars=c(2,3,1)), 1 )
                  grp known-DM unkn-DM pre-DM Well Survey
study  sex respl
DANHES M   2.0-3.0          7.7    10.0   12.3 14.5   14.0
           1.6-2.0          7.7    11.1   13.1 14.5   14.0
           1.3-1.5          7.7    12.2   13.5 14.4   14.0
           1.1-1.2          7.7    13.3   13.8 14.4   14.0
       F   2.0-3.0          7.8    10.0   12.2 14.4   14.0
           1.6-2.0          7.8    11.0   13.0 14.3   14.0
           1.3-1.5          7.8    12.1   13.4 14.3   14.0
           1.1-1.2          7.8    13.2   13.7 14.3   14.0
GESUS  M   2.0-3.0         33.9    37.2   40.5 43.9   42.7
           1.6-2.0         33.9    38.8   41.7 43.7   42.7
           1.3-1.5         33.9    40.3   42.3 43.6   42.7
           1.1-1.2         33.9    41.9   42.7 43.5   42.7
       F   2.0-3.0         30.9    35.2   39.5 43.8   42.7
           1.6-2.0         30.9    37.3   41.1 43.6   42.7
           1.3-1.5         30.9    39.3   41.8 43.5   42.7
           1.1-1.2         30.9    41.3   42.4 43.4   42.7
```

The flatter we make the shape the higher we see the response rates in the pre- and undiag-groups, but with deviations from the middle choice of (0,1,1.6,2.0) less than 1%, so with limited effect.

Hence we shall plot the resulting prevalence estimates for the middel choice (0,1,1.6,2.0).

3.3.1 Estimated prevalences

With the corrected results in hand we can now plot the age-specific prevalences for the 3 groups. We make two different kinds of plots; one with the age-speific prevalences of each of the three conditions, and one with the *stcked* prevalences by age.

```
> # library( devEMF )
> # postscript( "prv.eps", height=4, width=6 )
```

```
> # emf( "prv.emf", height=4, width=6 )
> # bmp( "prv.bmp", height=400, width=600 )
> plsep <-
+ function( st, sh="1.6-2.0" )
+     {
+ pra <- as.numeric( dimnames(prarr)[["age"]] )
+ par( mfrow=c(1,2), mar=c(3,0,1,1), oma=c(0,3,0,0), mgp=c(3,1,0)/1.6,
+      las=1, bty="n" )
+ clr <- c("red","limegreen","blue")
+ matplot( pra, cbind( prarr[sh,st,"M",,"Pop"   ,1:3],
+                      prarr[sh,st,"M",,"Survey",1:3] )*100,
+          type="l", lwd=3, lty=rep(c("solid","11"),each=3), col=clr, yaxs="i",
+          ylim=c(0,20), xlab="", lend="butt" )
+ text( 20, 19, "Men", adj=0 )
+ text( 20, 17, st, adj=0 )
+ text( 20, 15-0:2*1.2,c("DM","unknown DM","pre-DM"),col=clr, adj=0 )
+ matplot( pra, cbind( prarr[sh,st,"F",,"Pop"   ,1:3],
+                      prarr[sh,st,"F",,"Survey",1:3] )*100,
+          type="l", lwd=3, lty=rep(c("solid","11"),each=3), col=clr, yaxs="i",
+          ylim=c(0,20), xlab="", yaxt="n", lend="butt" )
+ text( 20,19, "Women", adj=0 )
+ # text( 20,17-0:2*1.2,c("DM","unknown DM","pre-DM"),col=clr, adj=0 )
+ mtext( "Prevalence (%)", side=2, outer=TRUE, las=0, line=1.6 )
+    }
> # dev.off()


> plsep("DANHES","1.6-2.0")


> plsep("GESUS","1.6-2.0")
```

We shall base conclusions only on the DANHES and the GESUS surveys, although the DANHES survey with the extremely low participation rate is less reliable.

```
> plsep("GESUS","2.0-3.0")


> plsep("GESUS","1.3-1.5")


> plsep("GESUS","1.1-1.2")
```

3.3.2 Distribution of DM states in the population

Finally, we make stacked plots of the prevalences of DM, unknown DM, pre DM and no DM for these two surveys, and for the chosen shape.

```
> # library( devEMF )
> # emf( "prv.emf", height=4, width=6 )
> # postscript( "prv.eps", height=400, width=600 )
> # bmp( "prv.bmp", height=400, width=600 )
> plstack <-
+ function( st, sh="1.6-2.0" )
+     {
+ pra <- as.numeric( dimnames(prarr)[["age"]] )
```

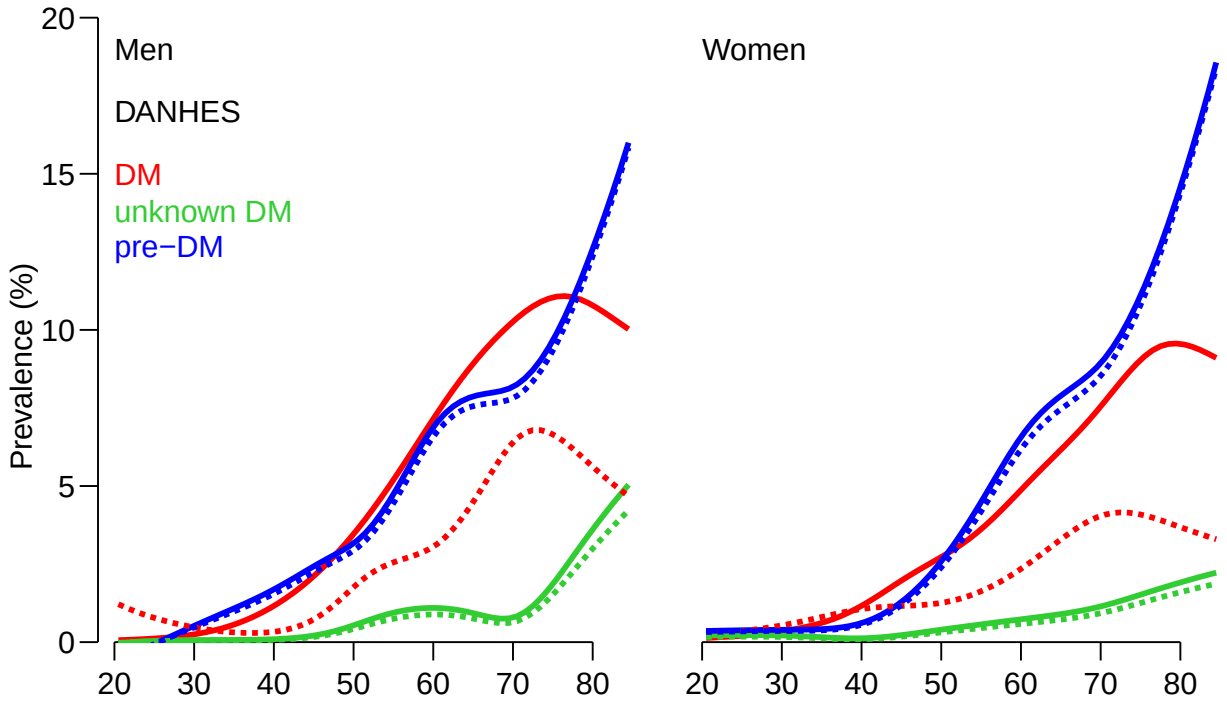

Figure 3.1: *Estimated age-specific prevalences of DM, uknown DM and pre-diabetes ine men (left) and women (right) in Denmark in 2011, based on the DANHES study. Median date of survey is 2008.3.*
Full lines are estimates of the population prevalences *and the broken lines are estimates of the* survey prevalences *(uncorrected) for the three groups. The population prevalences of DM are based on the register data at the median date of survey.* `../graph/DF-prDH`

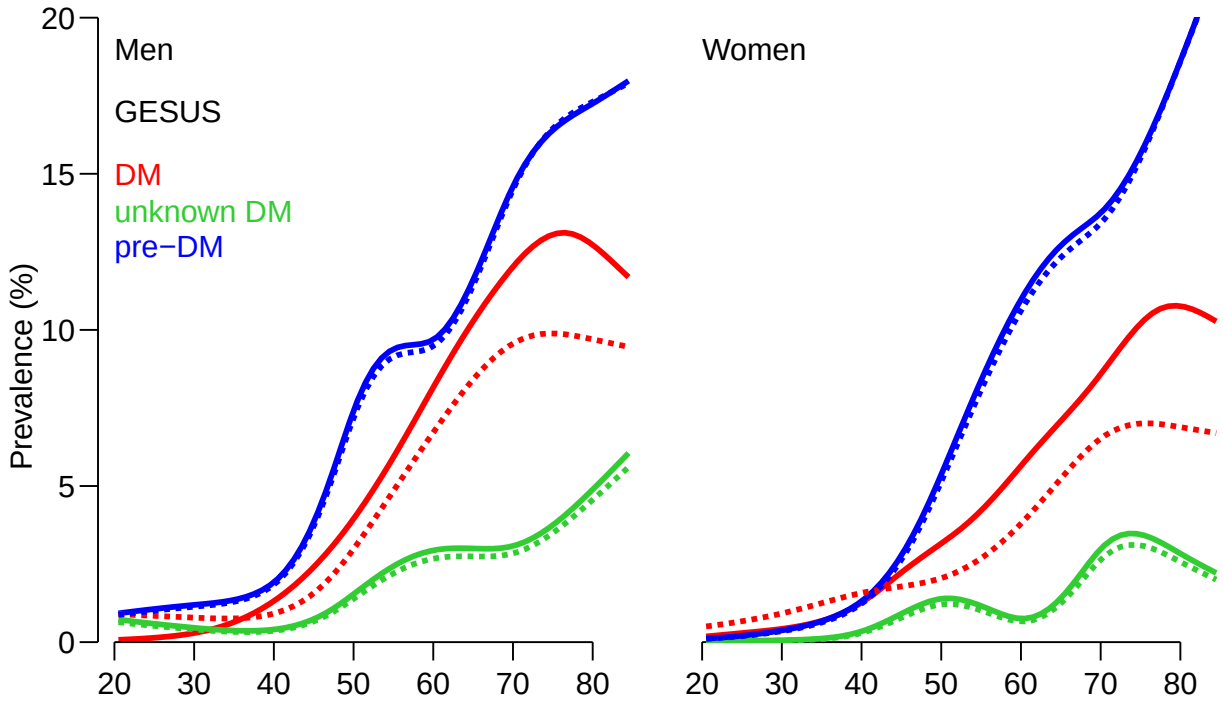

Figure 3.2: *Estimated age-specific prevalences of DM, uknown DM and pre-diabetes ine men (left) and women (right) in Denmark in 2011, based on the GESUS study. Median date of survey is 2011.5.*
Full lines are estimates of the population prevalences *and the broken lines are estimates of the* survey prevalences *(uncorrected) for the three groups. The population prevalences of DM are based on the register data at the median date of survey.* `../graph/DF-prGes`

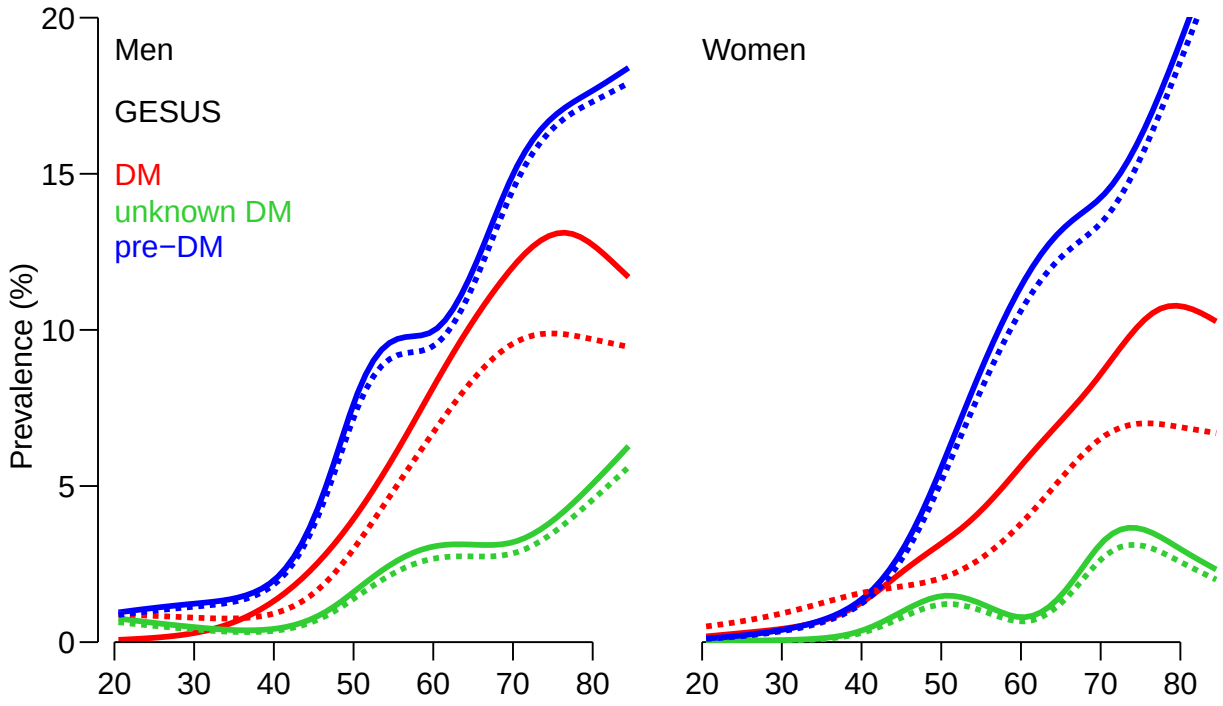

Figure 3.3: *Estimated age-specific prevalences of DM, uknown DM and pre-diabetes in men (left) and women (right) in Denmark in 2011, based on the GESUS study, using shape of responses (0,1,2,3). Median date of survey is 2012.*
Full lines are estimates of the population prevalences *and the broken lines are estimates of the* survey prevalences *(uncorrected) for the three groups. The population prevalences of DM are based on the register data at the median date of survey.* `../graph/DF-prGes20`

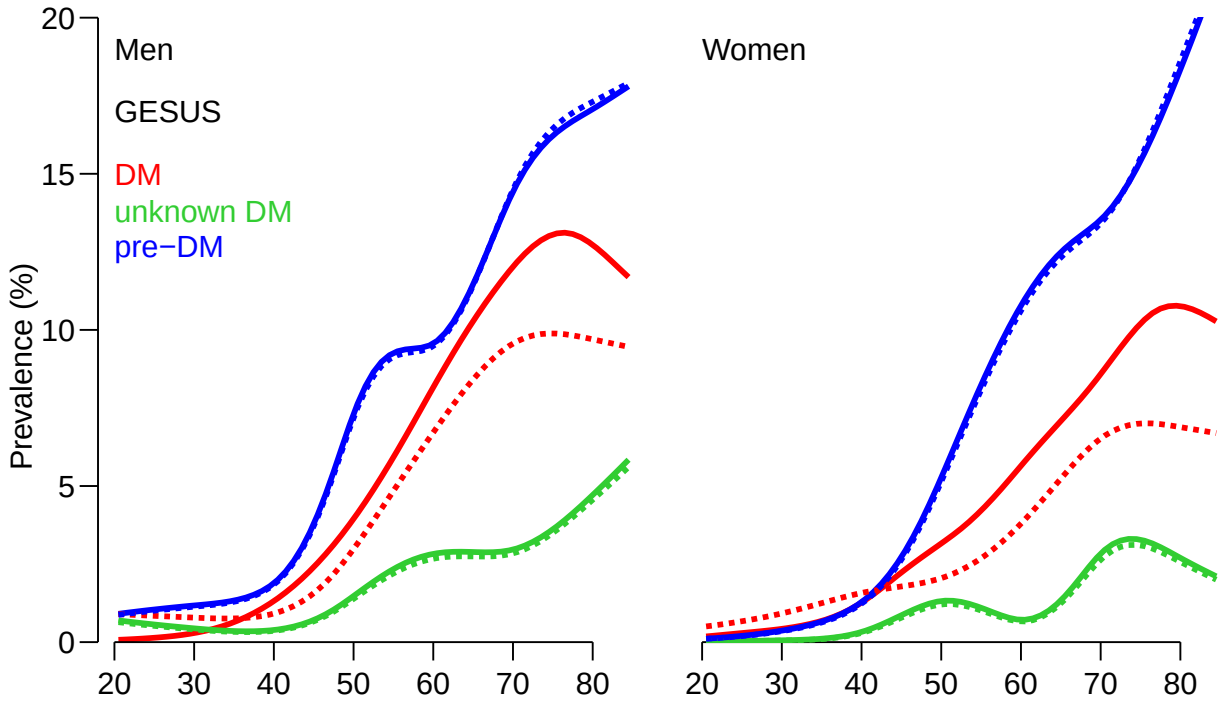

Figure 3.4: *Estimated age-specific prevalences of DM, uknown DM and pre-diabetes in men (left) and women (right) in Denmark in 2011, based on the GESUS study, using shape of responses (0,1,1.3,1.5). Median date of survey is 2011.5.*
Full lines are estimates of the population prevalences *and the broken lines are estimates of the* survey prevalences *(uncorrected) for the three groups. The population prevalences of DM are based on the register data at the median date of survey.* `../graph/DF-prGes13`

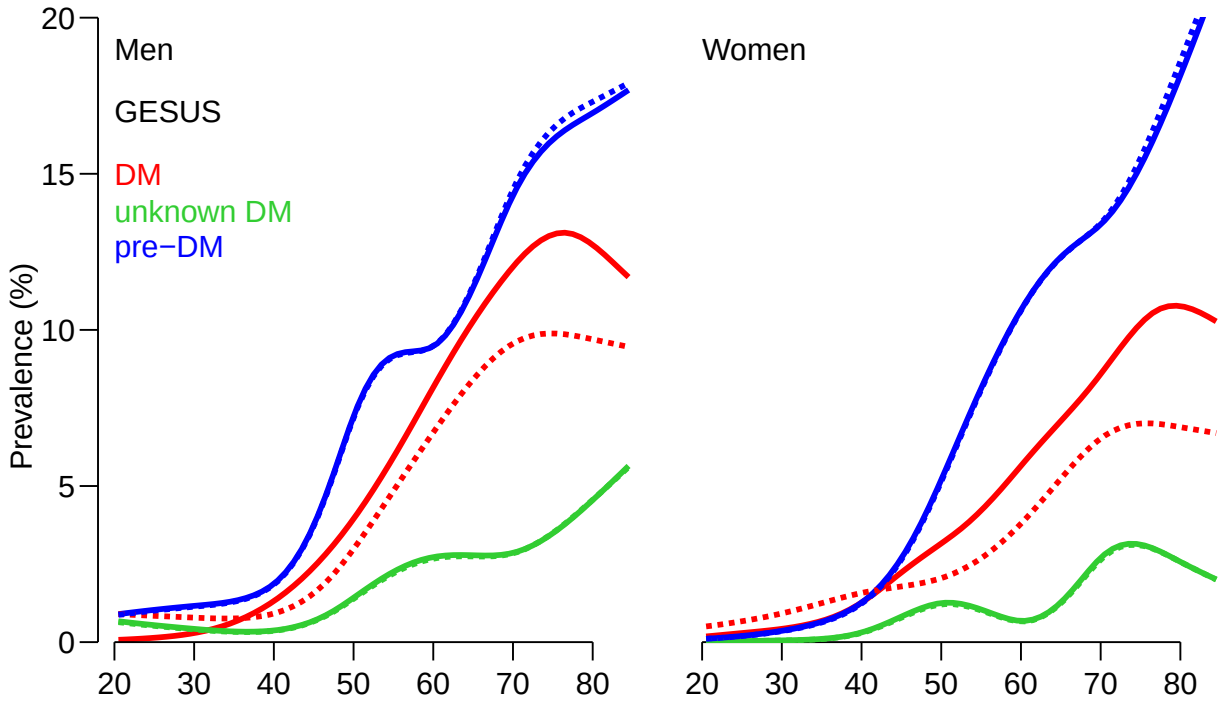

Figure 3.5: *Estimated age-specific prevalences of DM, uknown DM and pre-diabetes in men (left) and women (right) in Denmark in 2011, based on the GESUS study, using shape of responses (0,1,1.1,1.2). Median date of survey is 2011.5.*
Full lines are estimates of the population prevalences *and the broken lines are estimates of the* survey prevalences *(uncorrected) for the three groups. The population prevalences of DM are based on the register data at the median date of survey.* `../graph/DF-prGes11`

```
+ clr <- c("red","limegreen","lightblue","white")
+ cm <- cbind( 0, t( apply(prarr[sh,st,"M",,"Pop",],1,cumsum) ) )
+ cf <- cbind( 0, t( apply(prarr[sh,st,"F",,"Pop",],1,cumsum) ) )
+
+ par( mfrow=c(1,2), mar=c(2,0,1,1), oma=c(1,3,2,0), mgp=c(3,1,0)/1.6,
+      las=1, bty="n" )
+ plst <-
+ function(cm,sx)
+     {
+ plot( NA, xlab="", xaxs="i", xlim=c(20,85), xaxt="n",
+           ylab="", yaxs="i", ylim=c( 0,40), yaxt="n" )
+ axis( side=1, at=seq(30,80,10) )
+ axis( side=1, at=seq(20,85, 5), labels=NA, tcl=-0.3 )
+ if( sx=="Men" ) axis( side=2, at=0:8*5, labels=NA, tcl=-0.3 )
+ for ( i in 1:4 ) polygon( c(pra,rev(pra)),
+                           c(cm[,i],rev(cm[,i+1]))*100,
+                           col=clr[i], border=clr[i] )
+ abline( h=seq(5,95,5), col=gray(0.8), lty="21" )
+ text( 25, 39, sx, adj=0 )
+     }
+ plst( cm, "Men" )
+ axis( side=2 )
+ mtext( "Prevalence (%)", side=2, line=2, outer=TRUE, las=0 )
+ plst( cf, "Women" )
+ mtext( paste( st, "-study (", if( st=="DANHES") "03/2008)" else "05/2011)", sep="" ),
+        outer=TRUE, side=3, line=0 )
+ mtext( "Age", outer=TRUE, side=1, line=0 )
+   }


> plstack("DANHES")


> plstack("GESUS")
```

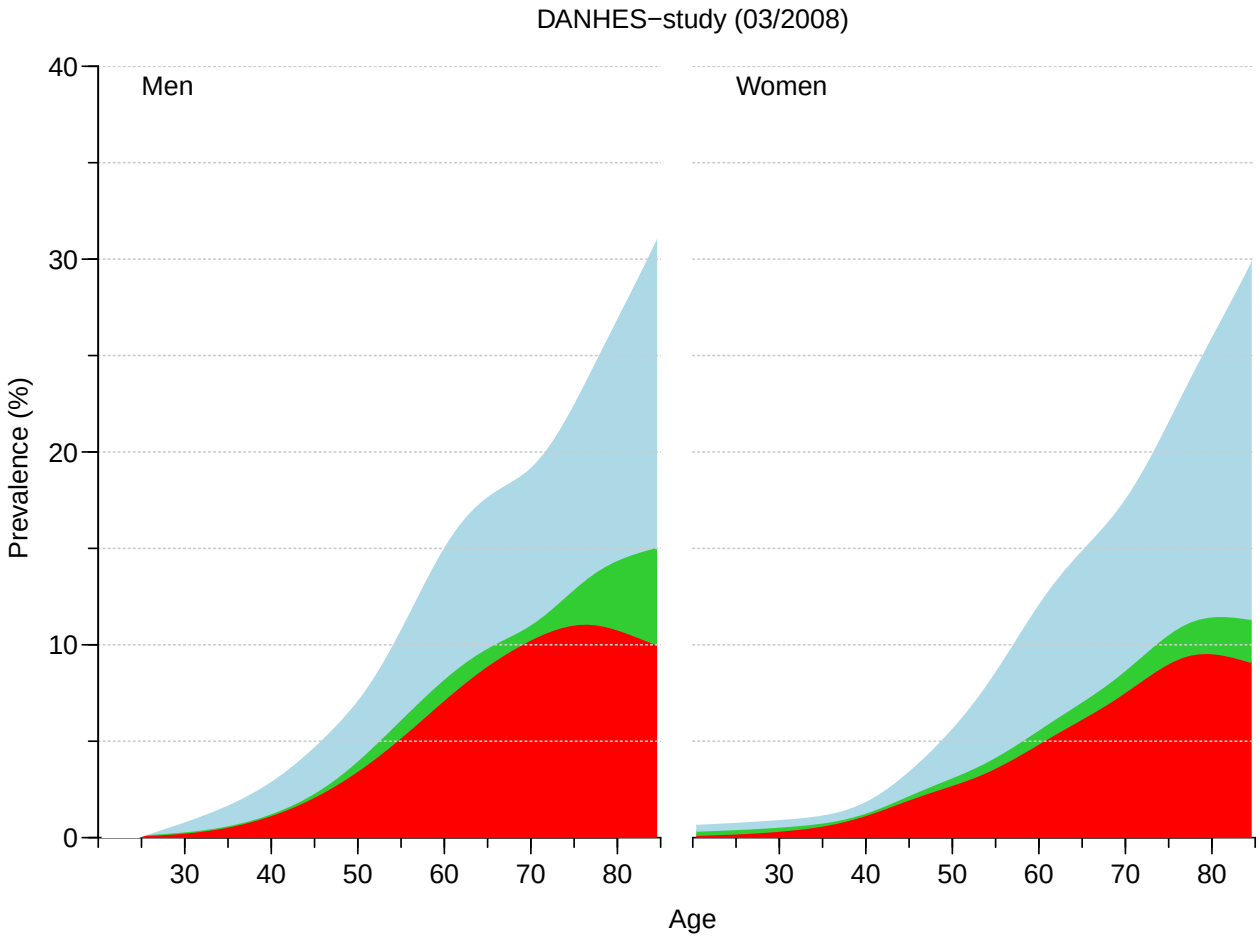

Figure 3.6: *Estimated age-specific prevalences of DM (red), uknown DM (green), pre-diabetes (light blue) and no diabetes (white) in men (left) and women (right) in Denmark in 2011, based on the DANHES study. Median date of survey is 2008.3.* `../graph/DF-stDH`

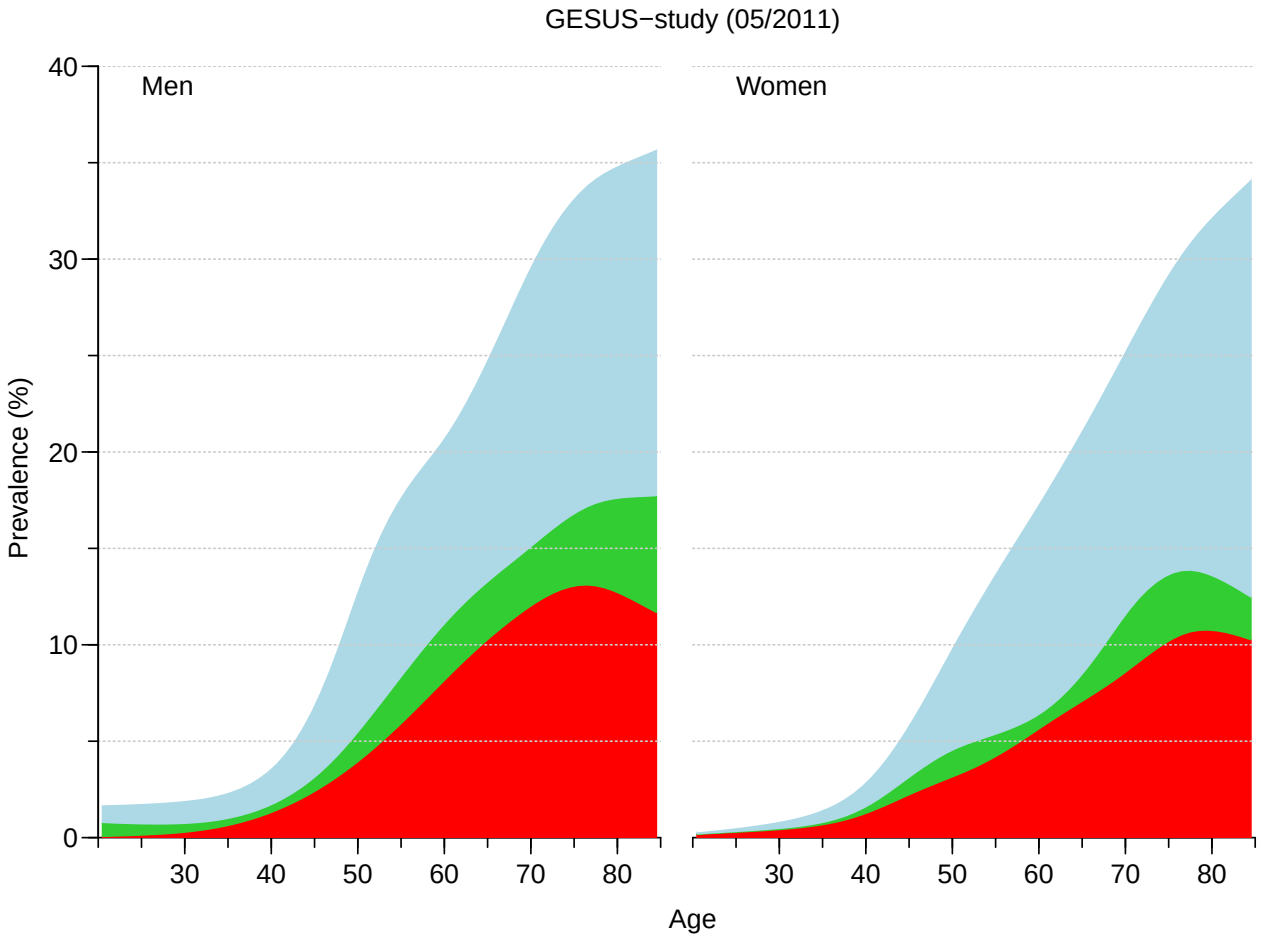

Figure 3.7: *Estimated age-specific prevalences of DM (red), uknown DM (green), pre-diabetes (light blue) and no diabetes (white) in men (left) and women (right) in Denmark in 2011, based on the GESUS study. Median date of survey is 2011.5.* `../graph/DF-stGes`

3.4 Number of persons with DM, unknown DM and pre-DM

In the array `prarr`, we have the age-specific prevalences of the four classes. Thus if we multiply these by the total population we get the *number* of persons with each condition.

We have the number of persons in the Danish population in the dataset `N.dk` (from the `Epi` package):

```
> data( N.dk )
> head( N.dk )
  sex A    P     N
1   1 0 1971 35839
2   2 0 1971 34108
3   1 1 1971 36302
4   2 1 1971 34153
5   1 2 1971 37855
6   2 2 1971 35609
```

Then we take the predicted prevalences of the four types of persons, and extend the array of predicted prevalences of DM to age 100, using the models for the empirical prevalences:

```
> str( prarr )
 num [1:4, 1:2, 1:2, 1:65, 1:2, 1:4] 0.01215 0.01215 0.01215 0.01215 0.00897 ...
 - attr(*, "dimnames")=List of 6
  ..$ respl: chr [1:4] "2.0-3.0" "1.6-2.0" "1.3-1.5" "1.1-1.2"
  ..$ study: chr [1:2] "DANHES" "GESUS"
  ..$ sex  : chr [1:2] "M" "F"
  ..$ age  : chr [1:65] "20.5" "21.5" "22.5" "23.5" ...
  ..$ type : chr [1:2] "Survey" "Pop"
  ..$ grp  : chr [1:4] "known-DM" "unkn-DM" "pre-DM" "Well"
> sh <- "1.6-2.0"
> Narr <- prarr[sh,c("DANHES","GESUS"),,,"Pop",]
> Narr <- Narr[,,c(1:65,rep(65,15)),]
> dimnames(Narr)[[3]] <- 20:99+0.5
> str( Narr )
 num [1:2, 1:2, 1:80, 1:4] 0.000628 0.000736 0.001334 0.00182 0.00072 ...
 - attr(*, "dimnames")=List of 4
  ..$ study: chr [1:2] "DANHES" "GESUS"
  ..$ sex  : chr [1:2] "M" "F"
  ..$ age  : chr [1:80] "20.5" "21.5" "22.5" "23.5" ...
  ..$ grp  : chr [1:4] "known-DM" "unkn-DM" "pre-DM" "Well"
```

We have now expanded the `Narr` with the ages 85 trough 99, and the `known-DM` category for these ages is then predicted from the models for the prevalences:

```
> ( mdate <- tapply( tot$doe, tot$st, median ) )
  DANHES     H-06     H-08    GESUS
2008.245 2007.500 2009.324 2011.599
> as.Date.cal.yr( mdate <- tapply( tot$doe, tot$st, median ) )
      DANHES         H-06         H-08        GESUS
"2008-03-31" "2007-07-03" "2009-04-29" "2011-08-08"
```

```
> nd.D <- data.frame( A=20:99+0.5, P=mdate["DANHES"] )
> nd.G <- data.frame( A=20:99+0.5, P=mdate["GESUS"] )
> Narr["DANHES","M",,"known-DM"] <- predict( mm, nd.D, type="response" )
> Narr["DANHES","F",,"known-DM"] <- predict( mw, nd.D, type="response" )
> Narr["GESUS" ,"M",,"known-DM"] <- predict( mm, nd.G, type="response" )
> Narr["GESUS" ,"F",,"known-DM"] <- predict( mw, nd.G, type="response" )
> str( Narr["GESUS",,1:65,"known-DM"] )
 num [1:2, 1:65] 0.000736 0.001831 0.00085 0.002076 0.00098 ...
 - attr(*, "dimnames")=List of 2
  ..$ sex: chr [1:2] "M" "F"
  ..$ age: chr [1:65] "20.5" "21.5" "22.5" "23.5" ...
> str( prarr[sh,"GESUS",,,"Pop","known-DM"] )
 num [1:2, 1:65] 0.000736 0.00182 0.00085 0.002061 0.00098 ...
 - attr(*, "dimnames")=List of 2
  ..$ sex: chr [1:2] "M" "F"
  ..$ age: chr [1:65] "20.5" "21.5" "22.5" "23.5" ...
> summary(as.vector( Narr["GESUS",,1:65,"known-DM"] / prarr[sh,"GESUS",,,"Pop","known-DM"] ))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.9997  0.9998  1.0007  1.0017  1.0031  1.0092
```

As a totally unfounded prediction into the blue, we let the prevalence of `unkn-DM` and `pre-DM` decay by age in the same pattern as `known-DM`, and then we adjust the `Well` category accordingly so that the sum is 1:

```
> Narr[,,66:80,2:3] <- Narr[,,66:80,c(1,1)]/Narr[,,rep(65,15),c(1,1)] *
+                                           Narr[,,rep(65,15),2:3]
> Narr[,,66:80,4] <- 1 - apply( Narr[,,66:80,1:3], 1:3, sum )
> summary( apply( Narr, 2:4, sum ) )
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.001367 0.027807 0.147675 0.500036 0.634910 1.989754
```

`Narr` now contains the probabilities (prevalences) of DM / unkn-DM / pre-DM / no-DM in ages 20–85, and a bold extrapoltion for ages 85–100.

In order to produce estimates of *numbers* of persons in each group at the median survey dates, we estimate the population sizes by sex and age by simple linear interpolation:

```
> Xarr <- xtabs( N ~ sex + A + P,
+                data = subset( N.dk, A>19 & A<100 & P %in% c(2008+0:1,2012+0:1) ) )
> str( Xarr )
 xtabs [1:2, 1:80, 1:4] 31662 30027 31720 30361 31140 ...
 - attr(*, "dimnames")=List of 3
  ..$ sex: chr [1:2] "1" "2"
  ..$ A  : chr [1:80] "20" "21" "22" "23" ...
  ..$ P  : chr [1:4] "2008" "2009" "2012" "2013"
 - attr(*, "call")= language xtabs(formula = N ~ sex + A + P, data = subset(N.dk, A > 19 & A <
> Parr <- Xarr[,,c(1,3)]
> # weighted population sizes to match survey median date
> Parr[,,1] <- (2009-dsarr["DANHES"]     ) * Xarr[,,"2008"] +
+              (     dsarr["DANHES"]-2008) * Xarr[,,"2009"]
> Parr[,,2] <- (2013-dsarr["GESUS"]      ) * Xarr[,,"2012"] +
+              (     dsarr["GESUS"] -2012) * Xarr[,,"2013"]
> str( Narr["GESUS",,,] )
```

```
 num [1:2, 1:80, 1:4] 0.000736 0.001831 0.00085 0.002076 0.00098 ...
 - attr(*, "dimnames")=List of 3
  ..$ sex: chr [1:2] "M" "F"
  ..$ age: chr [1:80] "20.5" "21.5" "22.5" "23.5" ...
  ..$ grp: chr [1:4] "known-DM" "unkn-DM" "pre-DM" "Well"
> str( Parr[,,rep("2012",4)] )
 table [1:2, 1:80, 1:4] 35181 34112 36542 35086 34882 ...
 - attr(*, "dimnames")=List of 3
  ..$ sex: chr [1:2] "1" "2"
  ..$ A  : chr [1:80] "20" "21" "22" "23" ...
  ..$ P  : chr [1:4] "2012" "2012" "2012" "2012"
> Narr["DANHES",,,] <- Narr["DANHES",,,]*Parr[,,rep("2008",4)]
> Narr["GESUS" ,,,] <- Narr["GESUS" ,,,]*Parr[,,rep("2012",4)]
> str( Narr )
 num [1:2, 1:2, 1:80, 1:4] 20.2 25.9 40.5 62.5 23.1 ...
 - attr(*, "dimnames")=List of 4
  ..$ study: chr [1:2] "DANHES" "GESUS"
  ..$ sex  : chr [1:2] "M" "F"
  ..$ age  : chr [1:80] "20.5" "21.5" "22.5" "23.5" ...
  ..$ grp  : chr [1:4] "known-DM" "unkn-DM" "pre-DM" "Well"
```

3.4.1 Surveyed age-range

For the semi-nice printing of tables we need a small formatting function formatting the nubers according to Danish conventions:

```
> dfm <- function(x,w=10,d=0) formatC(x,format="f",width=w,digits=d,big.mark=".",decimal.mark=
> efm <- function(x,w=10,d=0) formatC(x,format="f",width=w,digits=d,big.mark=",",decimal.mark=
```

First we summarize the number of persons in the different groups for the age-range 20–85:

```
> Ngr <- addmargins( apply( Narr[,,1:65,],c(1,2,4),sum), 2:3 )
> ftable( dfm(round(Ngr) ) )
           grp   known-DM    unkn-DM     pre-DM       Well        Sum
study  sex
DANHES M           81.030     12.603     78.190  1.824.120  1.995.942
       F           67.907     11.604     83.673  1.872.816  2.035.999
       Sum        148.936     24.206    161.862  3.696.936  4.031.941
GESUS  M           98.115     34.765    135.692  1.765.450  2.034.022
       F           80.867     22.350    135.567  1.834.449  2.073.234
       Sum        178.982     57.115    271.260  3.599.899  4.107.256
```

These are the number of persons in the surveyed age-range 20–85, and we can of course also compute the *overall* prevalence (in %) of these conditions in this age-range:

```
> ftable( dfm(sweep( Ngr, 1:2, Ngr[,,5], "/" )*100,w=5,d=1), 1 )
           grp known-DM unkn-DM pre-DM  Well   Sum
study  sex
DANHES M           4,1     0,6    3,9  91,4 100,0
       F           3,3     0,6    4,1  92,0 100,0
       Sum         3,7     0,6    4,0  91,7 100,0
GESUS  M           4,8     1,7    6,7  86,8 100,0
       F           3,9     1,1    6,5  88,5 100,0
       Sum         4,4     1,4    6,6  87,6 100,0
```

All ages

We can do the same, using the entire age-range 20–99:

```
> Ngr <- addmargins( apply( Narr,c(1,2,4),sum), 2:3 )
> ftable( dfm(round(Ngr) ) )
           grp   known-DM    unkn-DM     pre-DM       Well        Sum
study  sex
DANHES M           84.083     14.131     83.040  1.846.910  2.028.163
       F           74.294     13.165     96.704  1.926.861  2.111.026
       Sum        158.377     27.297    179.744  3.773.771  4.139.188
GESUS  M           101.923     36.738    141.551  1.788.860  2.069.073
       F           88.275     23.943    151.163  1.887.123  2.150.505
       Sum        190.199     60.681    292.715  3.675.983  4.219.577
```

These are the number of persons in the age-range 20–99, and we can of course also compute the *overall* prevalence (in %) of these conditions in this age-range:

```
> ftable( dfm(Pgr <- sweep( Ngr, 1:2, Ngr[,,5], "/" )*100,w=5,d=1), 1 )
           grp known-DM unkn-DM pre-DM  Well   Sum
study  sex
DANHES M           4,1     0,7    4,1  91,1 100,0
       F           3,5     0,6    4,6  91,3 100,0
       Sum         3,8     0,7    4,3  91,2 100,0
GESUS  M           4,9     1,8    6,8  86,5 100,0
       F           4,1     1,1    7,0  87,8 100,0
       Sum         4,5     1,4    6,9  87,1 100,0
```

For the poster table:

```
> str( Ngr )
 num [1:2, 1:3, 1:5] 84083 101923 74294 88275 158377 ...
 - attr(*, "dimnames")=List of 3
  ..$ study: chr [1:2] "DANHES" "GESUS"
  ..$ sex  : chr [1:3] "M" "F" "Sum"
  ..$ grp  : chr [1:5] "known-DM" "unkn-DM" "pre-DM" "Well" ...
> str( Pgr )
 num [1:2, 1:3, 1:5] 4.15 4.93 3.52 4.1 3.83 ...
 - attr(*, "dimnames")=List of 3
  ..$ study: chr [1:2] "DANHES" "GESUS"
  ..$ sex  : chr [1:3] "M" "F" "Sum"
  ..$ grp  : chr [1:5] "known-DM" "unkn-DM" "pre-DM" "Well" ...
> zz <- Ngr[,,c(1,1,2,2,3,3)]
> str( zz )
 num [1:2, 1:3, 1:6] 84083 101923 74294 88275 158377 ...
 - attr(*, "dimnames")=List of 3
  ..$ study: chr [1:2] "DANHES" "GESUS"
  ..$ sex  : chr [1:3] "M" "F" "Sum"
  ..$ grp  : chr [1:6] "known-DM" "known-DM" "unkn-DM" "unkn-DM" ...
> zz[,,c(2,4,6)] <- Pgr[,,1:3]
> round( ftable(zz), 1 )
           grp known-DM known-DM  unkn-DM  unkn-DM   pre-DM   pre-DM
study  sex
DANHES M        84082.7      4.1  14131.1      0.7  83039.5      4.1
       F        74294.5      3.5  13165.5      0.6  96704.2      4.6
       Sum     158377.2      3.8  27296.6      0.7 179743.7      4.3
GESUS  M       101923.4      4.9  36738.1      1.8 141551.2      6.8
       F        88275.4      4.1  23943.0      1.1 151163.4      7.0
       Sum     190198.8      4.5  60681.1      1.4 292714.6      6.9
```