Pre-diabetes and unknown diabetes in Denmark

SDCC / Diabetesforeningen
Juli 2018
http://bendixcarstensen.com/SDC/DF
Version 17

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

Dansk sammenfatning

1.1 Antal personer med diabetes, udiagnosticert diabetes og prædiabetes i Danmark

Formålet med denne rapport er at give et kvalificeret estimat af hvor mange personer i Danmark der har kendt diabetes, uerkendt diabetes samt prædiabetes.

1.2 Registre og surveys

Antallet af personer med diabetes kan estimeres præcist ud fra eksisterende registre, mens gruppen af personer med uerkendt diabetes hhv. prædiabetes må estimeres ud fra surveys. Det er velkendt at personers tilbøjeliged til at deltage i surveys afhænger af deres sygdomsstatus; generelt er raske mennesker mere tilbøjelige til at deltage i surveys. Således vil en ukritisk brug af surveys underestimere antallet af personer med diabetes, uerkendt diabetes og prædiabetes og overestimere antallet af personer uden.

Da vi har adgang til et diabetes register[1] kan vi sige noget om underrapporteringen i surveys hvor diabetes indgår som et udfald. En stor del af denne rapport beskriver i detaljer hvordan dette (kan) gøres, og dokumenterer hvad vi har gjort.

1.3 Data grundlag

Vi beskriver analyser af forekomsten af prædiabetes og uerkendt diabetes i Danmark, baseret på studierne KRAM (2007–2008) og Gesus (2010–2014). Vi har haft andre studier til rådighed også, men disse er vurderet som værende for gamle (Inter-99) eller for små og usikre (Helbred 2006, Helbred 2008).

En oversigt over alder og dato ved deltagelse i survey findes i figur 3.1 på side 23, for alle 5 surveys vi har haft til rådighed.

Det skal bemærkes at disse surveys kun omfatter aldersspændet 20–85 år, og de tal vi præsenterer her er derfor kun for disse aldre.

[1] Et diabetes register dannet i forbindelse med et forskningsprojekt — dannet ud fra andre sundhedsregistre på Danmarks Statistik. Indeholder også en klassifikation af paeronser som T1/T2. En detaljeret (dynamisk) beskrivelse af registeret kan findes som http://bendixcarstensen.com/DMreg/NewReg.pdf.

1.3.1 Definitioner

Surveys er fra 2013 og tidligere, så HbA_{1c} er målt i de "gamle" enheder, dvs %.

I alle surveys er deltagerne blevet spurgt om de har diabetes, men derudover er målt HbA_{1c} på deltagerne. Vi har brugt resultaterne af disse målinger til at definere uerkendt diabtes som en HbA_{1c}-måling over 6.4% og prædiabetes som en HbA_{1c}-måling mellem 6.0 og 6.4% (inkl.).

1.4 Resultater

Det estimerede antal personer med T2 diabetes, uerkendt diabetes, prædiabetes og uden diabetes i hele befolkningen mellem 20 og 85 år, baseret på de to surveys fra hhv. 2007–8 (KRAM; median dato matrs 2008) og 2010–14 (Gesus; dian dato maj 2011) findes i tabel 1.1 og andelen (%) i de fire grupper i tabel 1.2.

Resultater fra surveys er korrigeret for forskellige responsrater i de forskellige grupper. Kendt T2 diabetes er baseret på det reviderede diabetes register.

Tabel 1.1: *Estimeret antal personer i alderen 20–85 (hhv. 20–100) i* hele befolkningen *med diabetes, uerkendt diabetes, prædiabetes hhv. uden diabetes. Den mediane survey dato er for KRAM 2008-03-13 og for GESUS 2011-10-05 — tallene refererer til disse datoer.*

		kendt DM	udiag. DM	præDM	uden DM	hele bef.
Personer 20–85 år						
KRAM	Mænd	81.030	12.603	78.190	1.824.120	1.995.942
	Kvinder	67.907	11.604	83.673	1.872.816	2.035.999
	M + K	148.936	24.206	161.862	3.696.936	4.031.941
Gesus	Mænd	98.115	34.765	135.692	1.765.450	2.034.022
	Kvinder	80.867	22.350	135.567	1.834.449	2.073.234
	M + K	178.982	57.115	271.260	3.599.899	4.107.256
Personer 20–100 år						
KRAM	Mænd	84.083	14.131	83.040	1.846.910	2.028.163
	Kvinder	74.294	13.165	96.704	1.926.861	2.111.026
	M + K	158.377	27.297	179.744	3.773.771	4.139.188
Gesus	Mænd	101.923	36.738	141.551	1.788.860	2.069.073
	Kvinder	88.275	23.943	151.163	1.887.123	2.150.505
	M + K	190.199	60.681	292.715	3.675.983	4.219.577

Detaljerede tegninger af de aldersspecifikke forekomster er i figur 4.6 og 4.7 på siderne 44 og 45.

1.4.1 Personer over 85

Der er ikke nogen af surveys der har data for antallet af personer over 85 år med DM, ukendt DM hhv. prædiabetes.

Tabel 1.2: *Estimeret andel (%) personer i alderen 20–85 hhv. 20–100 i* hele befolkningen *med diabetes, uerkendt diabetes, prædiabetes hhv. uden diabetes. Den mediane survey dato er for KRAM 2008-03-13 og for GESUS 2011-10-05, — tallene refererer til disse datoer.*

		kendt DM	udiag. DM	præDM	uden DM	hele bef.
Personer 20–85 år						
KRAM	Mænd	4,1	0,6	3,9	91,4	100
	Kvinder	3,3	0,6	4,1	92,0	100
	M + K	3,7	0,6	4,0	91,7	100
Gesus	Mænd	4,8	1,7	6,7	86,8	100
	Kvinder	3,9	1,1	6,5	88,5	100
	M + K	4,4	1,4	6,6	87,6	100
Alle personer 20–100 år						
KRAM	Mænd	4,1	0,7	4,1	91,1	100
	Kvinder	3,5	0,6	4,6	91,3	100
	M + K	3,8	0,7	4,3	91,2	100
Gesus	Mænd	4,9	1,8	6,8	86,5	100
	Kvinder	4,1	1,1	7,0	87,8	100
	M + K	4,5	1,4	6,9	87,1	100

Personer over 85 udgør 1.3% af alle mænd og 2.7% af alle kvinder, hhv. 5.8% af mænd og 11.0% af kvinder over 60 år hvor de fleste diabetes tilfælde forekommer. Det er således ikke helt ligegyldigt for det samlede billede hvordan antallet af prædiabetikere og udiagnosticerede ser ud i denne aldersklasse. I mangel af pålidelige data her vi derfor foretaget en (alders)fremskrivning af prævalensen af prædiabetes og ukendt diabetes wed at antage den samme relative ændring (med alderen) af de aldersspecifikke prævalenser som vi ser i prævalensen af diabetes efter alder 85.

2

Background and theory

2.1 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.

2.2 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 2.1

```
> pi <- c(12,88)
> rr <- c(40,60)
> pp <- pi*rr/100
> pp <- pp / sum(pp) * 100
> par(mar=c(1,1,1,1),mgp=c(3,1,0)/1.6)
> plot( NA, axes=FALSE, xlim=0:1*100, ylim=0:1*100, xaxs="i", yaxs="i" )
> rect( 0, 0, pi[1],   100, col=gray(1.0) )
> rect( 0, 0, pi[1], rr[1], col=gray(0.5) )
> rect( pi[1], 0, 100,   100, col=gray(1.0) )
> rect( pi[1], 0, 100, rr[2], col=gray(0.5) )
> abline( h = sum(pi*rr/100), col="red ")
> text( cumsum(pi)-pi/2, rr/2, formatC(pp,format="f",digits=1), col="white" )
> text( cumsum(pi)-pi/2, 95  , formatC(pi,format="f",digits=1) )
> text( cumsum(pi)-pi/2, rr+2, formatC(rr,format="f",digits=1), adj=c(0.5,0) )
```

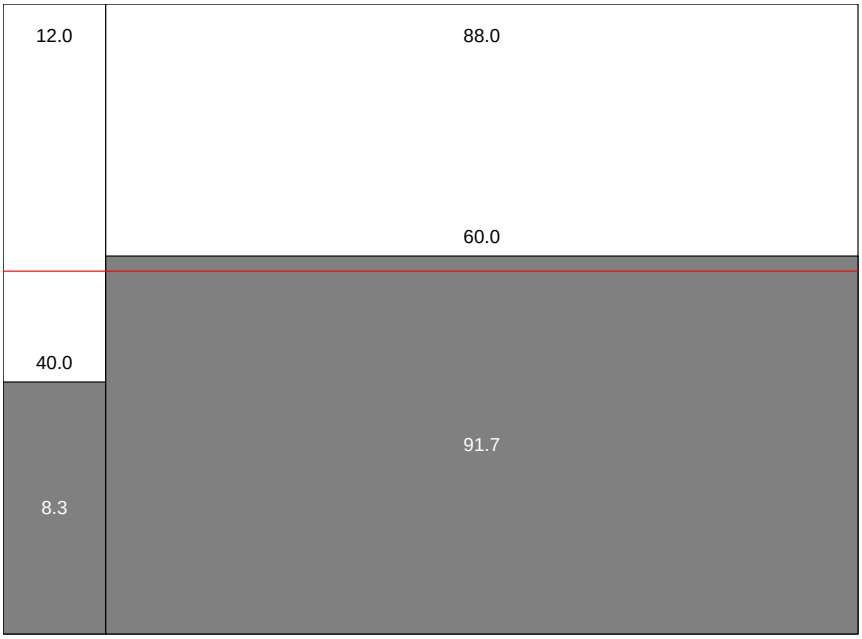

Figure 2.1: *An illustration of how differential response rates influence the result of a survey; 12% prevalence in the population, transforms to 8.3% prevalence in the surveyed part of the population. The red line is the overall survey response rate* `../graph/DF-simple`

2.2.1 General set-up

Taking the numerical illustration in figure 2.1 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
  - 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:

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

2.2.2 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.

2.3 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:

```
> pp <- c(11,13,17,59)/100
> pr <- c(30,40,50,55)/100
> res <- cbind( pp*(1-pr), pp*pr ) * 1000
> colnames( res ) <- c("noRep","Reply")
> rownames( res ) <- c("DM","unkDM","preDM","noDM")
> cbind( tt <- addmargins( res <- round( res ) ),
+        round( tt[,2]/tt[,3]*100, 1),
+        round( 100*sweep(tt,2,tt[5,],"/"), 1 ) )
      noRep Reply Sum      noRep Reply Sum
DM       77    33 110 30.0  15.2   6.7  11
unkDM    78    52 130 40.0  15.4  10.5  13
preDM    85    85 170 50.0  16.8  17.2  17
noDM    265   324 589 55.0  52.5  65.6  59
Sum     505   494 999 49.4 100.0 100.0 100
```

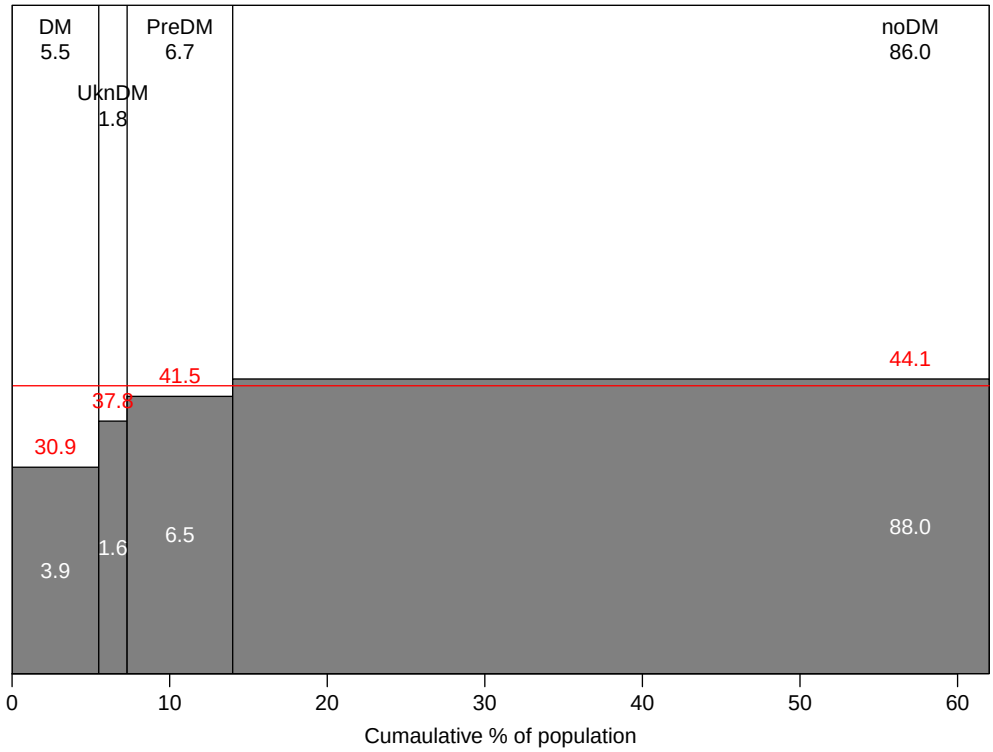

Figure 2.2: *Illustration of the result of a survey with differing response rates in different DM classes; numbers are precentages, taken to resemble those derived for men from the GESUS survey. Note that the x-axis does not extend to 100%. At the top is the population prevalences, corresponding to the* 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 red numbers are the response rates in each category. The gray area above the red line is the same as the white area below the red line.* `../graph/DF-groups`

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 2.2:

```
> xmx <- 62
> pi <- c(5.5,1.8,6.7,86.0) # c(11,13,17,59)
> rr <- c(30.9,37.8,41.5,44.1) # c(30,40,50,55)
> pp <- pi*rr/100
> pp <- pp / sum(pp) * 100
> par(mar=c(3,1,1,1),mgp=c(3,1,0)/1.6)
> plot( NA, axes=FALSE, xlim=0:1*xmx, ylim=0:1*100,
+       xaxs="i", yaxs="i", xlab="Cumaulative % of population" )
> axis( side=1 )
> rect( 0, 0, pi[1],   100, col=gray(1.0) )
> rect( 0, 0, pi[1], rr[1], col=gray(0.5) )
> # rect( pi[1], 0, xmx,   100, col=gray(1.0) )
> rect( pi[1], 0, xmx, rr[2], col=gray(0.5) )
> rect( sum(pi[1:2]), 0, xmx,   200, col=gray(1.0) )
> rect( sum(pi[1:2]), 0, xmx, rr[3], col=gray(0.5) )
> rect( sum(pi[1:3]), 0, xmx,   100, col=gray(1.0) )
> rect( sum(pi[1:3]), 0, xmx, rr[4], col=gray(0.5) )
```

```
> abline( h = sum(pi*rr)/100, col="red")
> text( cumsum(pi)-pi/2, rr/2, formatC(pp,format="f",digits=1), col="white" )
> text( cumsum(pi)-pi/2, c(95,85,95,95), paste( c("DM","UknDM","PreDM","noDM"), #rownames(res)
+                              formatC(pi,format="f",digits=1), sep="\n") )
> text( cumsum(pi)-pi/2, rr+2, formatC(rr,format="f",digits=1),
+      adj=c(0.5,0), col="red" )
> box(col="white",lwd=2)
> box(col="black",bty="c")
```

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 2.2. 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 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.

2.3.1 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:

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

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 2.2, 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.

2.4 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.

3

Data acquisition

3.1 Reading

We load the relevant R-packages for reading and processing data:

```
> library( haven )
> library( Epi )
> print( sessionInfo(), l=F )
R version 3.4.1 (2017-06-30)
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_2.19    haven_1.1.0

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.12      lattice_0.20-35   zoo_1.8-0         MASS_7.3-47
 [5] grid_3.4.1        plyr_1.8.4        magrittr_1.5      etm_0.6-2
 [9] rlang_0.1.1       Matrix_1.2-10     splines_3.4.1     forcats_0.2.0
[13] cmprsk_2.2-7      numDeriv_2016.8-1 survival_2.41-3   parallel_3.4.1
[17] compiler_3.4.1    tibble_1.3.3
```

3.2 Reading survey data with `haven`

First we read all the raw data from all studies (the SAS datasets with the `haven::read_sas` function; and the the GESUS with |read.csv|):

```
> i99 <- read_sas( "../data/inter99.sas7bdat" )
> add <- read_sas( "../data/addition.sas7bdat" )
> h8  <- read_sas( "../data/h8.sas7bdat" )
> h6  <- read_sas( "../data/h6.sas7bdat" )
> krm <- read_sas( "../data/kram.sas7bdat" )
> ges <- read.csv( "../data/gesus.csv", header=TRUE )
```

3.2.1 GESUS

The HbA_{1c}-values in GESUS are in mmol/mol, so we convert them to the "old units" in order to be able to use them across studies for study if pre-diabetes prevalence.

```
> str( ges )
'data.frame':        21205 obs. of  12 variables:
 $ visitid   : int  20834 20835 20836 20837 20838 20839 20840 20841 20842 20843 ...
 $ sex       : int  0 1 0 0 1 0 0 0 1 0 ...
 $ bdate     : Factor w/ 12654 levels "10/10/1926","10/10/1928",..: 6327 4244 6336 4225 5212 6
 $ usdate    : Factor w/ 601 levels "10/10/2011","10/10/2012",..: 58 58 58 58 58 58 58 58 58 5
 $ hba1c     : num  38.8 35.5 39.9 43.2 39.9 ...
 $ gluc      : num  5.2 5.6 5.4 5.7 5.8 6 5.5 5.1 5.1 4.7 ...
 $ m_20_15   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ dmjanej   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ l_96_33   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ insulin   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ l_96_35   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ andendmmed: int  0 0 0 0 0 0 0 0 0 0 ...
```

```
> head( ges )
  visitid sex    bdate    usdate    hba1c gluc m_20_15 dmjanej l_96_33 insulin l_96_35
1   20834   0 4/1/1937 1/11/2010 38.79781  5.2       0       0       0       0       0
2   20835   1 2/1/1977 1/11/2010 35.51913  5.6       0       0       0       0       0
3   20836   0 4/1/1946 1/11/2010 39.89071  5.4       0       0       0       0       0
4   20837   0 2/1/1953 1/11/2010 43.16940  5.7       0       0       0       0       0
5   20838   1 3/1/1934 1/11/2010 39.89071  5.8       0       0       0       0       0
6   20839   0 4/1/1949 1/11/2010 36.61202  6.0       0       0       0       0       0
  andendmmed
1          0
2          0
3          0
4          0
5          0
6          0
```

```
> names( ges )
 [1] "visitid"    "sex"        "bdate"      "usdate"     "hba1c"      "gluc"
 [7] "m_20_15"    "dmjanej"    "l_96_33"    "insulin"    "l_96_35"    "andendmmed"
```

```
> gesus <- data.frame( dob = as.Date(ges$bdate,format="%m/%d/%Y" ),
+                      doe = as.Date(ges$usdate,format="%m/%d/%Y" ),
+                      sex = factor( ges$sex,
+                                    levels=1:0,
+                                    labels=c("M","F") ),
+ # From mmol/mol to % (New number/10.929)+2.15
+ # From % to mmol/mol (Old number-2.15)x10.929
+                      hba = ges$hba1c / 10.929 + 2.15, # Now in %
+                       dm = factor( ges$dmjanej,
+                                    levels=1:0,
+                                    labels=c("Y","N") ),
+                       st = "GESUS" )
> gesus <- cal.yr( gesus )
> str( gesus )
```

```
'data.frame':        21205 obs. of  6 variables:
 $ dob:Classes 'cal.yr', 'numeric'  num [1:21205] 1937 1977 1946 1953 1934 ...
 $ doe:Classes 'cal.yr', 'numeric'  num [1:21205] 2010 2010 2010 2010 2010 ...
 $ sex: Factor w/ 2 levels "M","F": 2 1 2 2 1 2 2 2 1 2 ...
```

```
 $ hba: num  5.7 5.4 5.8 6.1 5.8 ...
 $ dm : Factor w/ 2 levels "Y","N": 2 2 2 2 2 2 2 2 2 2 ...
 $ st : Factor w/ 1 level "GESUS": 1 1 1 1 1 1 1 1 1 1 ...
```

```
> gesus <- gesus[complete.cases(gesus),]
> summary( gesus )
```

```
      dob            doe       sex            hba         dm            st
 Min.   :1914   Min.   :2010   M: 9532   Min.   : 2.498   Y:  930   GESUS:20941
 1st Qu.:1945   1st Qu.:2011   F:11409   1st Qu.: 5.352   N:20011
 Median :1955   Median :2012             Median : 5.535
 Mean   :1956   Mean   :2012             Mean   : 5.607
 3rd Qu.:1966   3rd Qu.:2013             3rd Qu.: 5.810
 Max.   :1990   Max.   :2014             Max.   :13.770
```

3.2.2 Inter 99

```
> names( i99 ) <- tolower( names(i99) )
> str( i99 )
```

```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':       6784 obs. of  6 variables:
 $ sex    : atomic  1 1 1 1 1 2 2 1 1 2 ...
  ..- attr(*, "label")= chr "Køn"
  ..- attr(*, "format.sas")= chr "F"
 $ fdato  : atomic  1.17e+10 1.17e+10 1.19e+10 1.20e+10 1.19e+10 ...
  ..- attr(*, "label")= chr "fødedato"
  ..- attr(*, "format.sas")= chr "F"
 $ scrdato: atomic  1.31e+10 1.31e+10 1.31e+10 1.32e+10 1.31e+10 ...
  ..- attr(*, "label")= chr "undersøgelsesdato (screeningsstation)"
  ..- attr(*, "format.sas")= chr "F"
 $ hba1c  : atomic  6.3 6.2 6.4 5.7 6.3 5.7 5.6 5.3 6.1 5.5 ...
  ..- attr(*, "label")= chr "B(Hgb) - Hæmoglobin A1C, stoffraktion"
  ..- attr(*, "format.sas")= chr "F"
 $ ald_udr: atomic  45.2 45.3 40.3 36.5 40.4 ...
  ..- attr(*, "label")= chr "alder ved us., udregnet [år]"
  ..- attr(*, "format.sas")= chr "F"
 $ kendtdm: atomic  0 0 0 0 0 0 0 0 0 0 ...
  ..- attr(*, "label")= chr "kendtdm"
 - attr(*, "label")= chr "INTER99"
```

```
> head( i99 )
```

```
# A tibble: 6 x 6
    sex       fdato     scrdato hba1c  ald_udr kendtdm
  <dbl>       <dbl>       <dbl> <dbl>    <dbl>   <dbl>
1     1 11714457600 13140835200   6.3 45.20014       0
2     1 11714457600 13143254400   6.2 45.27681       0
3     1 11872224000 13142736000   6.4 40.26096       0
4     1 12029990400 13180406400   5.7 36.45526       0
5     1 11872224000 13145846400   6.3 40.35952       0
6     2 12029990400 13143427200   5.7 35.28344       0
```

```
> head( (i99$scrdato-i99$fdato)/i99$ald_udr )
```

```
[1] 31556926 31556926 31556926 31556926 31556926 31556926
```

```
> ( spy <- 365.2425*24*60*60 ) # seconds per year
```

```
[1] 31556952
```

```
> i99[,c("fdato","scrdato")] <- i99[,c("fdato","scrdato")]/spy +
+                               cal.yr( as.Date("1582-10-14") )
>                               # start of Gregorian calendar
> head( i99 )
```

```
# A tibble: 6 x 6
    sex    fdato  scrdato hba1c  ald_udr kendtdm
  <dbl>    <dbl>    <dbl> <dbl>    <dbl>   <dbl>
1     1 1954.008 1999.208   6.3 45.20014       0
2     1 1954.008 1999.284   6.2 45.27681       0
3     1 1959.007 1999.268   6.4 40.26096       0
4     1 1964.006 2000.462   5.7 36.45526       0
5     1 1959.007 1999.367   6.3 40.35952       0
6     2 1964.006 1999.290   5.7 35.28344       0
```

However, it is well known that the Inter-99 has inflated HbA_{1c}-values, by 0.5% [**?**], so this is corrected here:

```
> int99 <- data.frame( dob = i99$fdato,
+                      doe = i99$scrdato,
+                      sex = factor( i99$sex,
+                                    levels=1:2,
+                                    labels=c("M","F") ),
+                      hba = i99$hba1c - 0.5,
+                       dm = factor( i99$kendtdm,
+                                    levels=1:0,
+                                    labels=c("Y","N") ),
+                       st = "I-99" )
> int99 <- zap_formats( int99 )
> str( int99 )
```

```
'data.frame':        6784 obs. of  6 variables:
 $ dob: atomic  1954 1954 1959 1964 1959 ...
  ..- attr(*, "label")= chr "fødedato"
 $ doe: atomic  1999 1999 1999 2000 1999 ...
  ..- attr(*, "label")= chr "undersøgelsesdato (screeningsstation)"
 $ sex: Factor w/ 2 levels "M","F": 1 1 1 1 1 2 2 1 1 2 ...
 $ hba: atomic  5.8 5.7 5.9 5.2 5.8 5.2 5.1 4.8 5.6 5 ...
  ..- attr(*, "label")= chr "B(Hgb) - Hæmoglobin A1C, stoffraktion"
 $ dm : Factor w/ 2 levels "Y","N": 2 2 2 2 2 2 2 2 2 2 ...
 $ st : Factor w/ 1 level "I-99": 1 1 1 1 1 1 1 1 1 1 ...
```

3.2.3 DANHES

The DANHES study records grommed into the relevant format:

```
> names( krm ) <- tolower( names(krm) )
> str( krm )
```

```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':        18065 obs. of  13 variables:
 $ fdate       : atomic  -5478 -3287 1096 3288 -6543 ...
  ..- attr(*, "label")= chr "Fødselsdato"
 $ aldvintr_bus: atomic  62 56 44 38 65 57 50 60 49 40 ...
  ..- attr(*, "label")= chr "Alder"
 $ dato        : Date, format: "2007-03-28" "2007-04-13" ...
 $ hgba1c      : atomic  NA 5.2 5.6 5.6 8.9 5.1 7.3 5.6 5.6 5.2 ...
  ..- attr(*, "label")= chr "HgBA1c"
  ..- attr(*, "format.sas")= chr "COMMA"
```

```
 $ gender_bus  : atomic  2 2 2 1 2 1 1 2 2 1 ...
  ..- attr(*, "label")= chr "Køn"
  ..- attr(*, "format.sas")= chr "GENDERF"
 $ helbredvgt1 : atomic  0.91 0.91 0.891 1.406 0.91 ...
  ..- attr(*, "label")= chr "Opvægter helbredsundersøgelsen til stikprøven - vægtene har snit
 $ s73         : atomic  2 2 2 2 1 2 1 2 2 NA ...
  ..- attr(*, "label")= chr "73. Har du nogensinde fået at vide af en læge, at du havde sukker
  ..- attr(*, "format.sas")= chr "S73LAB"
 $ s74         : atomic  NA NA NA NA 53 NA 33 NA NA NA ...
  ..- attr(*, "label")= chr "74. Hvor gammel var du, da en læge fortalte dig, at du havde sukk
  ..- attr(*, "format.sas")= chr "S10LAB"
 $ s75a        : atomic  NA NA NA NA 1 NA NA NA NA NA ...
  ..- attr(*, "label")= chr "75.a. Hvilken behandling får du nu for din sukkersyge/diabetes? a
  ..- attr(*, "format.sas")= chr "S75ALAB"
 $ s75b        : atomic  NA NA NA NA NA NA NA NA NA NA ...
  ..- attr(*, "label")= chr "75.b. Hvilken behandling får du nu for din sukkersyge/diabetes? b
  ..- attr(*, "format.sas")= chr "S75BLAB"
 $ s75c        : atomic  NA NA NA NA 1 NA NA NA NA NA ...
  ..- attr(*, "label")= chr "75.c. Hvilken behandling får du nu for din sukkersyge/diabetes? c
  ..- attr(*, "format.sas")= chr "S75CLAB"
 $ s75d        : atomic  NA NA NA NA 1 NA 1 NA NA NA ...
  ..- attr(*, "label")= chr "75.d. Hvilken behandling får du nu for din sukkersyge/diabetes? c
  ..- attr(*, "format.sas")= chr "S75DLAB"
 $ s75e        : atomic  NA NA NA NA NA NA NA NA NA NA ...
  ..- attr(*, "label")= chr "75.e. Hvilken behandling får du nu for din sukkersyge/diabetes? e
  ..- attr(*, "format.sas")= chr "S75ELAB"
 - attr(*, "label")= chr "KRAM"
```

```
> krm$fdate <- krm$fdate/365.25 + 1960
> krm$dato  <- cal.yr( krm$dato )
> head( krm )
```

```
# A tibble: 6 x 13
     fdate aldvintr_bus     dato hgba1c gender_bus helbredvgt1   s73   s74  s75a  s75b
     <dbl>        <dbl>    <dbl>  <dbl>      <dbl>       <dbl> <dbl> <dbl> <dbl> <dbl>
1 1945.002           62 2007.235     NA          2   0.9096990     2    NA    NA    NA
2 1951.001           56 2007.279    5.2          2   0.9097216     2    NA    NA    NA
3 1963.001           44 2007.287    5.6          2   0.8906308     2    NA    NA    NA
4 1969.002           38 2007.290    5.6          1   1.4064747     2    NA    NA    NA
5 1942.086           65 2007.270    8.9          2   0.9096579     1    53     1    NA
6 1950.086           57 2007.309    5.1          1   0.9081696     2    NA    NA    NA
# ... with 3 more variables: s75c <dbl>, s75d <dbl>, s75e <dbl>
```

```
> with( krm, table(floor(s74/10)*10,s73,useNA="ifany") )
```

```
      s73
           1     2  <NA>
  0        6     0     0
  10      22     0     0
  20      21     0     0
  30      30     0     1
  40      51     0     0
  50     107     0     0
  60     124     0     0
  70      38     0     0
  80       2     0     0
  <NA>    14 17052   597
```

```
> kram <- data.frame( dob = krm$fdate,
+                     doe = krm$dato,
+                     sex = factor( krm$gender_bus,
```

```
+                                    levels=1:2,
+                                    labels=c("M","F") ),
+                       hba = krm$hgba1c,
+                        dm = factor(krm$s73,
+                                    levels=1:2,
+                                    labels=c("Y","N") ),
+                        st = "DANHES" )
> kram <- zap_formats(kram)
> str( kram )
'data.frame':        18065 obs. of  6 variables:
 $ dob: atomic  1945 1951 1963 1969 1942 ...
  ..- attr(*, "label")= chr "Fødselsdato"
 $ doe:Classes 'cal.yr', 'numeric'  num [1:18065] 2007 2007 2007 2007 2007 ...
 $ sex: Factor w/ 2 levels "M","F": 2 2 2 1 2 1 1 2 2 1 ...
 $ hba: atomic  NA 5.2 5.6 5.6 8.9 5.1 7.3 5.6 5.6 5.2 ...
  ..- attr(*, "label")= chr "HgBA1c"
 $ dm : Factor w/ 2 levels "Y","N": 2 2 2 2 1 2 1 2 2 NA ...
 $ st : Factor w/ 1 level "DANHES": 1 1 1 1 1 1 1 1 1 1 ...
```

3.2.4 Helbred 2006

```
> names( h6 ) <- tolower( names( h6 ) )
> head( h6 )
# A tibble: 6 x 8
       foedt    regdato   sex fldkendtdiabetes  hb2c  hb2d hb2l1 fldhba1c
      <date>     <date> <dbl>             <dbl> <dbl> <dbl> <dbl>    <dbl>
1 1936-03-30 2007-03-07     2                 2     2    NA     2      5.9
2 1936-04-08 2007-03-01     2                 2     2    NA     2      4.9
3 1936-04-09 2006-09-07     2                NA     2    NA     2      5.9
4 1936-04-28 2008-01-08     1                 1     1     1     1      7.4
5 1936-04-30 2007-07-05     1                 2     2    NA     2      6.4
6 1936-05-09 2007-07-17     2                 2     2    NA     2      5.9
> str( h6 )
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':        3471 obs. of  8 variables:
 $ foedt           : Date, format: "1936-03-30" "1936-04-08" ...
 $ regdato         : Date, format: "2007-03-07" "2007-03-01" ...
 $ sex             : atomic  2 2 2 1 1 2 1 2 2 2 ...
  ..- attr(*, "label")= chr "Køn"
  ..- attr(*, "format.sas")= chr "KOEN"
 $ fldkendtdiabetes: atomic  2 2 NA 1 2 2 1 2 NA 2 ...
  ..- attr(*, "label")= chr "har deltager kendt diabetes"
  ..- attr(*, "format.sas")= chr "JANEJ"
 $ hb2c            : atomic  2 2 2 1 2 2 1 2 2 2 ...
  ..- attr(*, "label")= chr "2.C. Har du nogensinde fået at vide, at du har sukkersyge?"
  ..- attr(*, "format.sas")= chr "JANEJ"
 $ hb2d            : atomic  NA NA NA 1 NA NA 2 NA NA NA ...
  ..- attr(*, "label")= chr "2.D. Får du behandling for sukkersyge?"
  ..- attr(*, "format.sas")= chr "JANEJ"
 $ hb2l1           : atomic  2 2 2 1 2 2 1 2 2 2 ...
  ..- attr(*, "label")= chr "2.L. Har en læge nogensinde fortalt dig, at du havde/har1. Sukker
  ..- attr(*, "format.sas")= chr "JANEJ"
 $ fldhba1c        : atomic  5.9 4.9 5.9 7.4 6.4 ...
  ..- attr(*, "label")= chr "Hba1c, Glucosyleret hæmoglobin"
 - attr(*, "label")= chr "HELBRED2006_MARIT"
```

```
> head( h6 )
```

```
# A tibble: 6 x 8
       foedt    regdato   sex fldkendtdiabetes  hb2c  hb2d hb2l1 fldhba1c
      <date>     <date> <dbl>            <dbl> <dbl> <dbl> <dbl>    <dbl>
1 1936-03-30 2007-03-07     2                2     2    NA     2      5.9
2 1936-04-08 2007-03-01     2                2     2    NA     2      4.9
3 1936-04-09 2006-09-07     2               NA     2    NA     2      5.9
4 1936-04-28 2008-01-08     1                1     1     1     1      7.4
5 1936-04-30 2007-07-05     1                2     2    NA     2      6.4
6 1936-05-09 2007-07-17     2                2     2    NA     2      5.9
```

```
> helb6 <- data.frame( dob = h6$foedt,
+                      doe = h6$regdato,
+                      sex = factor( h6$sex,
+                                    levels=1:2,
+                                    labels=c("M","F") ),
+                      hba = h6$fldhba1c,
+                       dm = factor(h6$hb2c,
+                                   levels=1:2,
+                                   labels=c("Y","N") ),
+                       st = "H-06" )
> helb6 <- cal.yr( zap_formats(helb6) )
> str( helb6 )
```

```
'data.frame':        3471 obs. of  6 variables:
 $ dob:Classes 'cal.yr', 'numeric'  num [1:3471] 1936 1936 1936 1936 1936 ...
 $ doe:Classes 'cal.yr', 'numeric'  num [1:3471] 2007 2007 2007 2008 2008 ...
 $ sex: Factor w/ 2 levels "M","F": 2 2 2 1 1 2 1 2 2 2 ...
 $ hba: atomic  5.9 4.9 5.9 7.4 6.4 ...
  ..- attr(*, "label")= chr "Hba1c, Glucosyleret hæmoglobin"
 $ dm : Factor w/ 2 levels "Y","N": 2 2 2 1 2 2 1 2 2 2 ...
 $ st : Factor w/ 1 level "H-06": 1 1 1 1 1 1 1 1 1 1 ...
```

3.2.5 Helbred 2008

```
> names( h8 ) <- tolower( names(h8) )
> dim( h8 )
```

```
[1] 795   8
```

```
> str( h8 )
```

```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':        795 obs. of  8 variables:
 $ deltnr          : atomic  10003 10006 10008 10017 10018 ...
  ..- attr(*, "label")= chr "deltnr"
 $ fldkendtdiabetes: atomic  2 2 2 2 2 2 2 2 2 2 ...
  ..- attr(*, "label")= chr "har deltager kendt diabetes"
  ..- attr(*, "format.sas")= chr "JANEJ"
 $ regdato         : Date, format: "2008-11-04" "2008-10-07" ...
 $ sex             : atomic  1 1 1 2 2 2 2 2 2 2 ...
  ..- attr(*, "label")= chr "Køn"
  ..- attr(*, "format.sas")= chr "KOEN"
 $ foedt           : Date, format: "1955-12-19" "1961-02-05" ...
 $ hb2c            : atomic  2 2 2 2 2 2 2 2 2 2 ...
  ..- attr(*, "label")= chr "2.C. Har du nogensinde fået at vide, at du har sukkersyge?"
  ..- attr(*, "format.sas")= chr "JANEJ"
 $ hb2d            : atomic  NA NA NA NA NA NA NA NA NA NA ...
  ..- attr(*, "label")= chr "2.D. Får du behandling for sukkersyge?"
```

```
  ..- attr(*, "format.sas")= chr "JANEJ"
 $ fldhba1c        : atomic  5.1 5.5 5.2 5.2 5.6 ...
  ..- attr(*, "label")= chr "Hba1c, Glucosyleret hæmoglobin"
 - attr(*, "label")= chr "H8"
```

```
> head( h8 )
```

```
# A tibble: 6 x 8
  deltnr fldkendtdiabetes    regdato   sex      foedt  hb2c  hb2d fldhba1c
   <dbl>            <dbl>     <date> <dbl>     <date> <dbl> <dbl>    <dbl>
1  10003                2 2008-11-04     1 1955-12-19     2    NA      5.1
2  10006                2 2008-10-07     1 1961-02-05     2    NA      5.5
3  10008                2 2008-11-06     1 1954-09-30     2    NA      5.2
4  10017                2 2008-10-13     2 1966-12-02     2    NA      5.2
5  10018                2 2008-10-30     2 1955-11-14     2    NA      5.6
6  10021                2 2008-10-08     2 1973-11-15     2    NA      5.2
```

```
> helb8 <- data.frame( dob = h8$foedt,
+                      doe = h8$regdato,
+                      sex = factor( h8$sex,
+                                    levels=1:2,
+                                    labels=c("M","F") ),
+                      hba = h8$fldhba1c,
+                       dm = factor(h8$hb2c,
+                                   levels=1:2,
+                                   labels=c("Y","N") ),
+                       st = "H-08" )
> helb8 <- cal.yr( zap_formats(helb8) )
> str( helb8 )
```

```
'data.frame':        795 obs. of  6 variables:
 $ dob:Classes 'cal.yr', 'numeric'  num [1:795] 1956 1961 1955 1967 1956 ...
 $ doe:Classes 'cal.yr', 'numeric'  num [1:795] 2009 2009 2009 2009 2009 ...
 $ sex: Factor w/ 2 levels "M","F": 1 1 1 2 2 2 2 2 2 2 ...
 $ hba: atomic  5.1 5.5 5.2 5.2 5.6 ...
  ..- attr(*, "label")= chr "Hba1c, Glucosyleret hæmoglobin"
 $ dm : Factor w/ 2 levels "Y","N": 2 2 2 2 2 2 2 2 2 2 ...
 $ st : Factor w/ 1 level "H-08": 1 1 1 1 1 1 1 1 1 1 ...
```

3.2.6 Addition

We do not have a variable indicating the person's diabetes status because all of the persons are diabetic in this study:

```
> head( add )
```

```
# A tibble: 6 x 5
     id   base_dob base_sex base_date_0 base_hba1c
  <dbl>     <date>    <dbl>      <date>      <dbl>
1     1 1937-05-20        1  2001-04-04        5.5
2     2 1938-02-01        1  2001-04-03        5.8
3     3 1943-08-16        1  2001-04-03        5.3
4     4 1941-08-23        1  2001-04-02        5.9
5     5 1937-11-10        0  2001-04-02        5.8
6     6 1944-02-06        1  2001-04-02        5.4
```

Thus we do not include this study in the analyses and hence neither in the combined data set:

3.3 Combining the data sources

We now combine the data frames and define persons' diabetes status on the basis of their HbA_{1c}; using the following cuts for persons without known diabetes (values are only recorded to the nearest 0.1%):

- Well if $HbA_{1c} < 5.95$
- pre-DM if $5.95 < HbA_{1c} < 6.45$
- undiagnosed DM if $HbA_{1c} > 6.45$

The latter is only used for those without a known diagnosis of diabetes.

```
> tot <- rbind( int99, kram, helb6, helb8, gesus )
> tot$dmst <- factor( ifelse( tot$dm=="Y", 3, (tot$hba>5.95) + (tot$hba>6.45) ),
+                     levels = 3:0,
+                     labels = rev(c("Well","pre-DM","unkn-DM","known-DM")) )
> summary( tot )
      dob            doe        sex            hba             dm              st
 Min.   :1913   Min.   :1999   M:22093   Min.   : 2.498   Y   : 1631   I-99  : 6784
 1st Qu.:1946   1st Qu.:2008   F:27963   1st Qu.: 5.261   N   :47796   DANHES:18065
 Median :1955   Median :2009             Median : 5.444   NA's:  629   H-06  : 3471
 Mean   :1955   Mean   :2009             Mean   : 5.513                H-08  :  795
 3rd Qu.:1965   3rd Qu.:2011             3rd Qu.: 5.700                GESUS :20941
 Max.   :1991   Max.   :2014             Max.   :15.300
                                         NA's   :474

       dmst
 known-DM: 1631
 unkn-DM :  540
 pre-DM  : 3132
 Well    :43686
 NA's    : 1067
```

```
> with( tot, ftable( addmargins( table( sex, st, dm, dmst, useNA="ifany" ),
+                                margin = 1:2 ),
+                    row.vars = 1:3 ) )
              dmst known-DM unkn-DM pre-DM  Well    NA
sex st     dm
M   I-99   Y             72       0      0     0     0
           N              0      57    225  2946     2
           NA             0       0      0     0     0
    DANHES Y            209       0      0     0     0
           N              0      47    361  6297   173
           NA             0       0      0     0   273
    H-06   Y             77       0      0     0     0
           N              0      14     61  1382     5
           NA             0       0      0     0    14
    H-08   Y              0       0      0     0     0
           N              0       2     12   330     2
           NA             0       0      0     0     0
    GESUS  Y            521       0      0     0     0
           N              0     199    852  7960     0
           NA             0       0      0     0     0
    Sum    Y            879       0      0     0     0
           N              0     319   1511 18915   182
           NA             0       0      0     0   287
```

```
F    I-99    Y             67       0      0     0     0
             N              0      25    127  3257     6
             NA             0       0      0     0     0
     DANHES  Y            206       0      0     0     0
             N              0      50    451  9428   245
             NA             0       0      0     0   325
     H-06    Y             65       0      0     0     0
             N              0       9     81  1743     5
             NA             0       0      0     0    15
     H-08    Y              5       0      0     0     0
             N              0       1     24   417     0
             NA             0       0      0     0     2
     GESUS   Y            409       0      0     0     0
             N              0     136    938  9926     0
             NA             0       0      0     0     0
     Sum     Y            752       0      0     0     0
             N              0     221   1621 24771   256
             NA             0       0      0     0   342
Sum  I-99    Y            139       0      0     0     0
             N              0      82    352  6203     8
             NA             0       0      0     0     0
     DANHES  Y            415       0      0     0     0
             N              0      97    812 15725   418
             NA             0       0      0     0   598
     H-06    Y            142       0      0     0     0
             N              0      23    142  3125    10
             NA             0       0      0     0    29
     H-08    Y              5       0      0     0     0
             N              0       3     36   747     2
             NA             0       0      0     0     2
     GESUS   Y            930       0      0     0     0
             N              0     335   1790 17886     0
             NA             0       0      0     0     0
     Sum     Y           1631       0      0     0     0
             N              0     540   3132 43686   438
             NA             0       0      0     0   629
> 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;

```
> par( mar=c(3,3,1,1), mgp=c(3,1,0)/1.6, las=1, bty="n" )
> clr <- rainbow(nc<-nlevels(tot$st)) # c("blue",gray(0.6),"red","orange")
> with( tot, plot( doe, doe-dob,
+                  pch=".", col=clr[st], cex=0.4,
+                  xlab="Date of examination",
+                  ylab="Age at examination" ) )
> text( rep(2003,3), 92.5-1:nc*5, levels(tot$st), col=clr, font=1, adj=0, cex=1.0 )
> axis( side=1, at=1999:2014    , labels=NA, tcl=-0.3 )
> axis( side=2, at=seq(20,90, 5), labels=NA, tcl=-0.3 )
> axis( side=2, at=seq(20,90,10), labels=NA, tcl=-0.4 )
> abline( h=c(20,85), col=gray(0.9) )
```

It is clear from figure 3.1 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:

```
> resr <- c( 0.140, 0.447, 0.440, 0.427 )
> names( resr ) <- levels( tot$st )[-1]
> resr
DANHES   H-06   H-08  GESUS
 0.140  0.447  0.440  0.427
> save( tot, resr, file="../data/tot.Rda" )
```

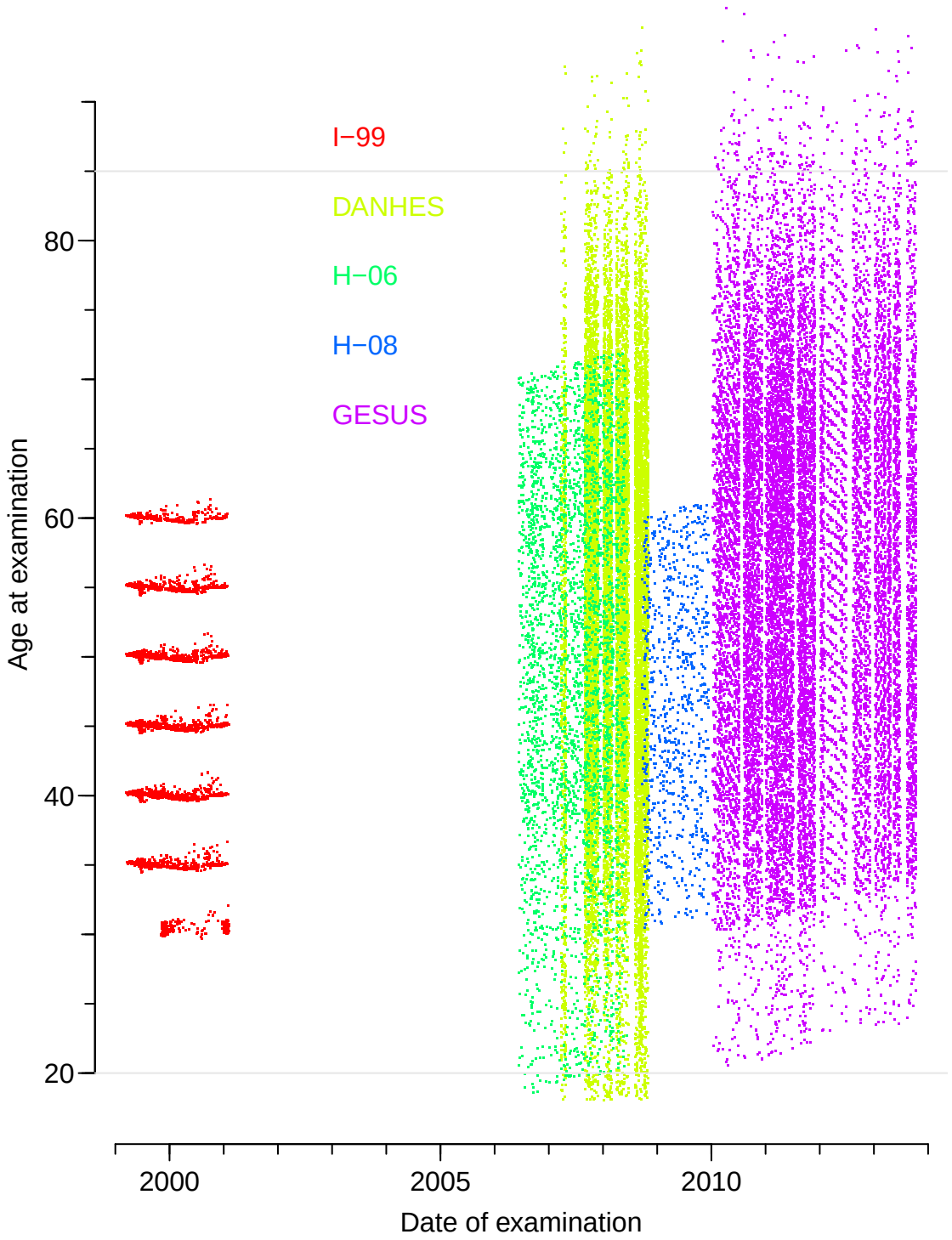

Figure 3.1: *Date and age at examination for the studies in this analysis.* `../graph/DF-exdat`

3.4 Register prevalence of T2D

We finally read the data from the reconstructed diabetes register:

```
> library( Epi )
> clear()
> load( "/home/bendix/sdc/DMreg/NewReg/data/DM1x1.Rda" )
> lls()
  name  mode class      dim          size(Kb)
1 FUtab list data.frame 4000 8          236.2
2 PRtab list data.frame 4400 7          225.2
> str( PRtab )
'data.frame':        4400 obs. of  7 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  1995 1995 1995 1995 1995 ...
 $ T1 : num  0 1 3 4 8 8 7 5 17 13 ...
 $ T2 : num  0 0 0 1 2 1 2 0 2 4 ...
 $ DM : num  0 1 3 5 10 9 9 5 19 17 ...
 $ N  : num  35821 34258 34790 33026 35059 ...
> prev <- PRtab[,c("sex","A","P","T2","N")]
> prev <- subset( prev, P>1995 )
> names( prev )[4] <- "X"
> str( prev )
'data.frame':        4200 obs. of  5 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  0 1 0 1 2 1 2 1 3 0 ...
 $ N  : num  36258 34198 36077 34452 35003 ...
> head( prev )
    sex A    P X     N
201   M 0 1996 0 36258
202   F 0 1996 1 34198
203   M 1 1996 0 36077
204   F 1 1996 1 34452
205   M 2 1996 2 35003
206   F 2 1996 1 33290
```

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") )
> summary( mm )

Call:
glm(formula = 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"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.2963  -0.9634  -0.1064   0.7653   4.3493

Coefficients: (1 not defined because of singularities)
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -8.74696    0.08757 -99.885  < 2e-16
Ns(A, kn = a.kn)1    2.71481    0.02615 103.820  < 2e-16
Ns(A, kn = a.kn)2    3.79862    0.02805 135.433  < 2e-16
Ns(A, kn = a.kn)3    4.61577    0.03624 127.370  < 2e-16
Ns(A, kn = a.kn)4    5.21366    0.04431 117.671  < 2e-16
Ns(A, kn = a.kn)5    5.57141    0.05333 104.462  < 2e-16
Ns(A, kn = a.kn)6    5.49487    0.06283  87.450  < 2e-16
Ns(A, kn = a.kn)7    6.87735    0.07032  97.802  < 2e-16
Ns(A, kn = a.kn)8    5.09034    0.07302  69.714  < 2e-16
Ns(P, kn = p.kn)1    0.48002    0.01243  38.614  < 2e-16
Ns(P, kn = p.kn)2    0.60670    0.01615  37.565  < 2e-16
Ns(P, kn = p.kn)3    1.00378    0.02776  36.163  < 2e-16
Ns(P, kn = p.kn)4    0.59469    0.02007  29.627  < 2e-16
Ns(P - A, kn = c.kn)1  0.07211    0.04246   1.698 0.089472
Ns(P - A, kn = c.kn)2  0.34649    0.05736   6.041 1.53e-09
Ns(P - A, kn = c.kn)3  0.41405    0.07141   5.798 6.70e-09
Ns(P - A, kn = c.kn)4  0.50565    0.08805   5.743 9.30e-09
Ns(P - A, kn = c.kn)5  0.54074    0.10242   5.280 1.29e-07
Ns(P - A, kn = c.kn)6  0.97267    0.12471   7.799 6.23e-15
Ns(P - A, kn = c.kn)7  0.26382    0.07929   3.327 0.000877
Ns(P - A, kn = c.kn)8       NA         NA      NA       NA

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2790384.2  on 2099  degrees of freedom
Residual deviance:    3263.3  on 2080  degrees of freedom
AIC: 17531

Number of Fisher Scoring iterations: 5
> summary( mw )

Call:
glm(formula = 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 == "F"))

Deviance Residuals:
   Min      1Q  Median      3Q     Max
```

```
-6.786  -1.309  -0.204   1.033   5.543

Coefficients: (1 not defined because of singularities)
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -9.66261    0.10157  -95.13   <2e-16
Ns(A, kn = a.kn)1     2.59373    0.02804   92.50   <2e-16
Ns(A, kn = a.kn)2     3.83363    0.03284  116.75   <2e-16
Ns(A, kn = a.kn)3     4.47670    0.04348  102.95   <2e-16
Ns(A, kn = a.kn)4     5.30049    0.05346   99.14   <2e-16
Ns(A, kn = a.kn)5     5.83796    0.06407   91.12   <2e-16
Ns(A, kn = a.kn)6     5.94344    0.07360   80.75   <2e-16
Ns(A, kn = a.kn)7     7.78314    0.08686   89.61   <2e-16
Ns(A, kn = a.kn)8     5.57060    0.08357   66.66   <2e-16
Ns(P, kn = p.kn)1     0.30698    0.01389   22.10   <2e-16
Ns(P, kn = p.kn)2     0.36553    0.01816   20.13   <2e-16
Ns(P, kn = p.kn)3     0.56771    0.03099   18.32   <2e-16
Ns(P, kn = p.kn)4     0.31370    0.02259   13.89   <2e-16
Ns(P - A, kn = c.kn)1 0.53503    0.04131   12.95   <2e-16
Ns(P - A, kn = c.kn)2 0.84498    0.05998   14.09   <2e-16
Ns(P - A, kn = c.kn)3 1.28898    0.07728   16.68   <2e-16
Ns(P - A, kn = c.kn)4 1.78103    0.09743   18.28   <2e-16
Ns(P - A, kn = c.kn)5 1.71233    0.11178   15.32   <2e-16
Ns(P - A, kn = c.kn)6 3.84084    0.14455   26.57   <2e-16
Ns(P - A, kn = c.kn)7 1.52799    0.07950   19.22   <2e-16
Ns(P - A, kn = c.kn)8       NA         NA      NA       NA

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2141214.5  on 2099  degrees of freedom
Residual deviance:    6266.2  on 2080  degrees of freedom
AIC: 20875

Number of Fisher Scoring iterations: 5
```

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:

```
> nd <- data.frame( A=rep(c(20:90,NA),5),
+                   P=rep(1996+0:4*5,each=72) )
> str( nd )
'data.frame':        360 obs. of  2 variables:
 $ A: int  20 21 22 23 24 25 26 27 28 29 ...
 $ P: num  1996 1996 1996 1996 1996 ...

> prM <- ci.pred( mm, nd )
> prW <- ci.pred( mw, nd )
> 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" )
> matplot( nd$A, prM*100,
+          type="l", lwd=c(2,1,1), lty=c(1,3,3), col="blue", yaxs="i",
+          ylim=c(0,20), xlab="" )
> text( 20,19, "Men", col="blue", adj=0 )
> matplot( nd$A, prW*100,
+          type="l", lwd=c(2,1,1), lty=c(1,3,3), col="red", yaxt="n", yaxs="i",
+          ylim=c(0,20), xlab="" )
> text( 20,19, "Women", col="red", adj=0 )
> mtext( "T2D prevalence 1996, 2001, 2006, 2011, 2016 (%)", side=2, outer=TRUE, las=0, line=1.
```

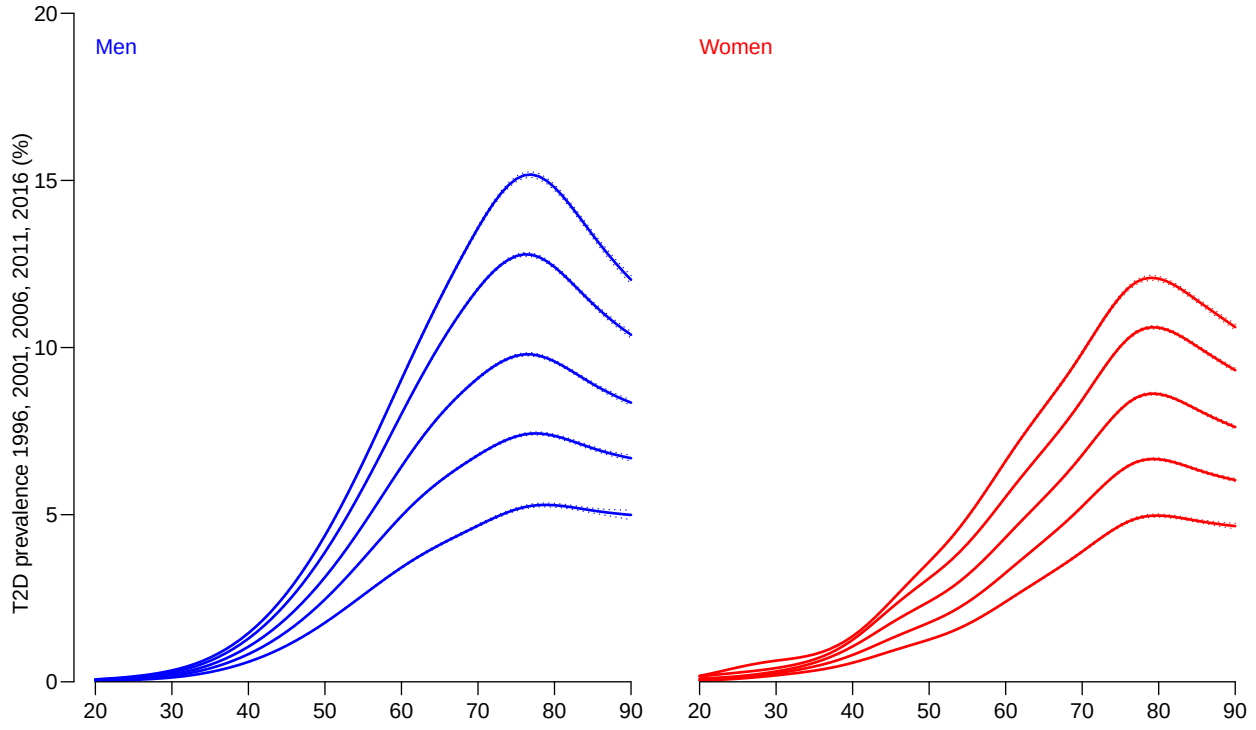

Figure 3.2: *Predicted prevalences of T2D from the reconstructed diabetes register (using an age-period-cohort model for the prevalences classified in 1-year age-classes) for each 1st January 1996, 2001, 2006, 2011, 2016, with 95% confidence limits.* `../graph/DF-prev`

Finally we save the model objects for further use:

```
> save( mm, mw, a.kn, p.kn, c.kn, file="../data/prmod.Rda" )
```

3.4.1 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
    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, w=10, d=0 )
```

```
          Ag  [-Inf,20)    [20,85)  [85, Inf)        Sum
sex P
```

```
1   2007              0    469,707     31,354    501,061
    2008              0    487,725     32,066    519,791
    2009              0    503,909     32,686    536,595
    2010              0    517,247     33,542    550,789
    2011              0    528,450     34,373    562,823
    2012              0    539,820     35,484    575,304
    2013              0    550,618     36,370    586,988
2   2007              0    540,243     74,501    614,744
    2008              0    554,648     74,778    629,426
    2009              0    567,942     75,770    643,712
    2010              0    578,820     76,421    655,241
    2011              0    588,318     76,943    665,261
    2012              0    598,733     77,557    676,290
    2013              0    607,891     78,143    686,034
```

```
> 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.

4

Analysis correcting for response

In this chapter, we implement the algorithm described in section 2.4.

4.1 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
```

4.2 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).

4.3 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).

4.3.1 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")
```

4.3.2 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"]] )
```

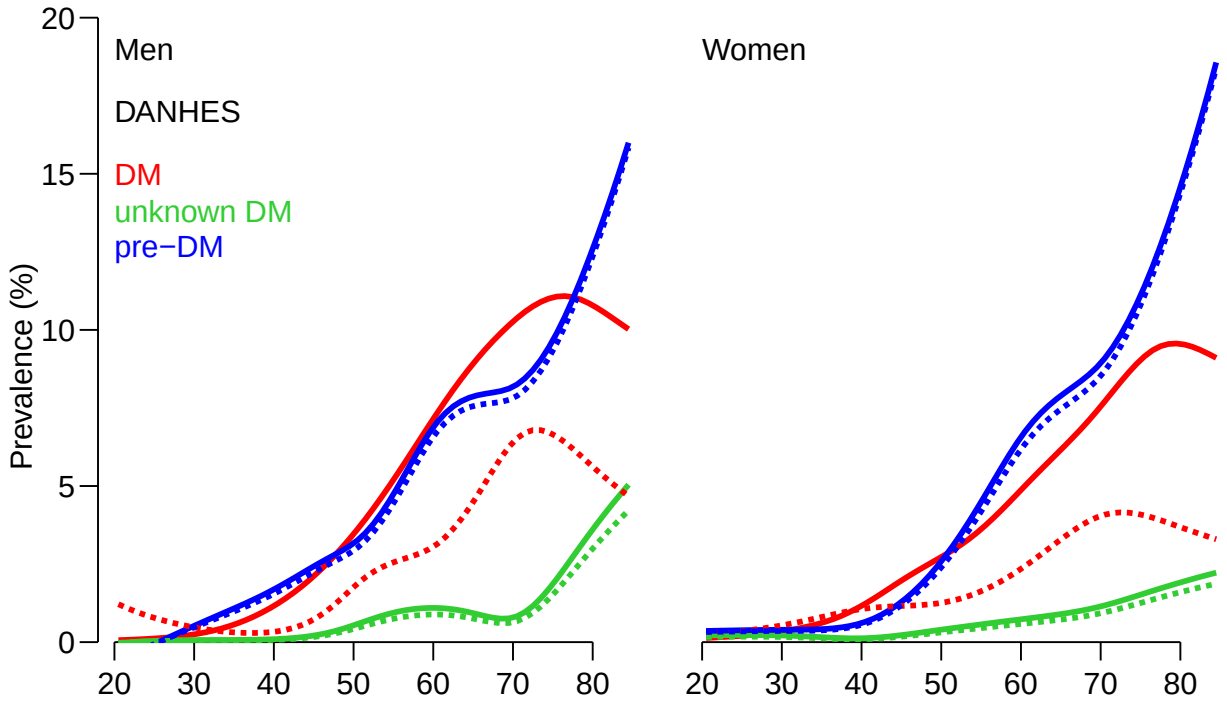

Figure 4.1: *Estimated age-specific prevalences of DM, uknown DM and pre-diabetes ine men (left) and women (right) in Denmark in 2011, based on the DANHES study. Median date of survey is 2008.3.*
Full lines are estimates of the population prevalences *and the broken lines are estimates of the* survey 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`

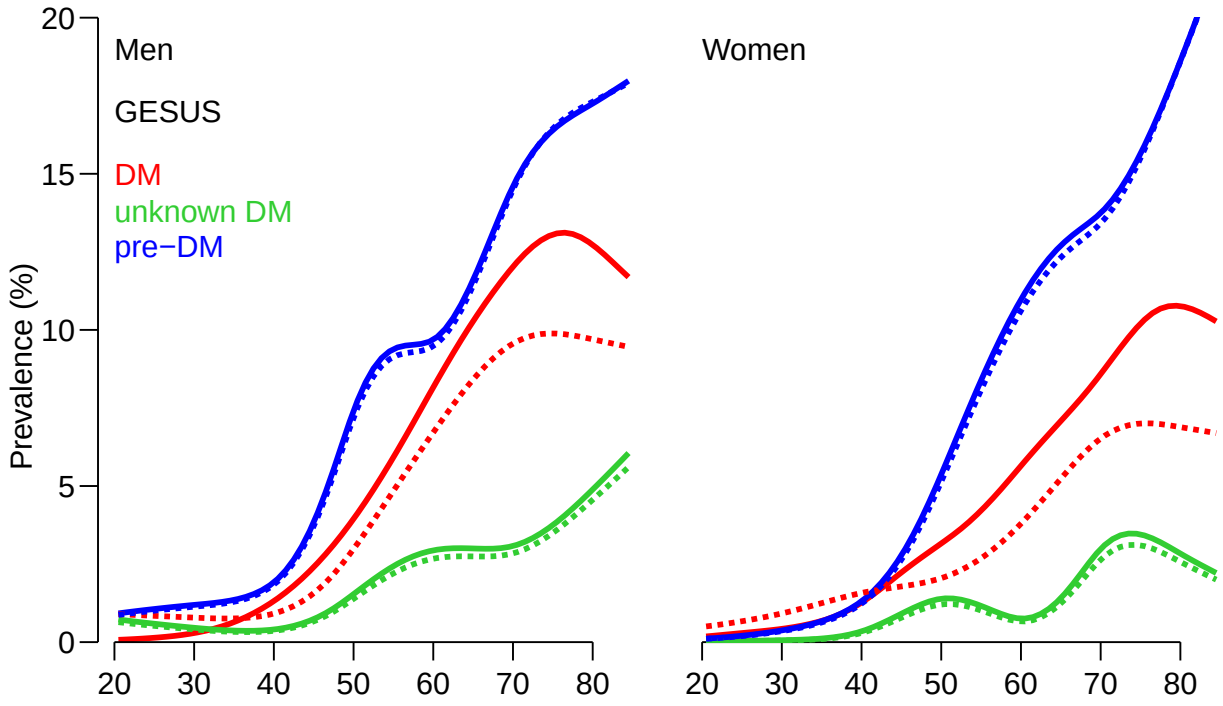

Figure 4.2: *Estimated age-specific prevalences of DM, uknown DM and pre-diabetes ine men (left) and women (right) in Denmark in 2011, based on the GESUS study. Median date of survey is 2011.5.*
Full lines are estimates of the population prevalences *and the broken lines are estimates of the* survey 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`

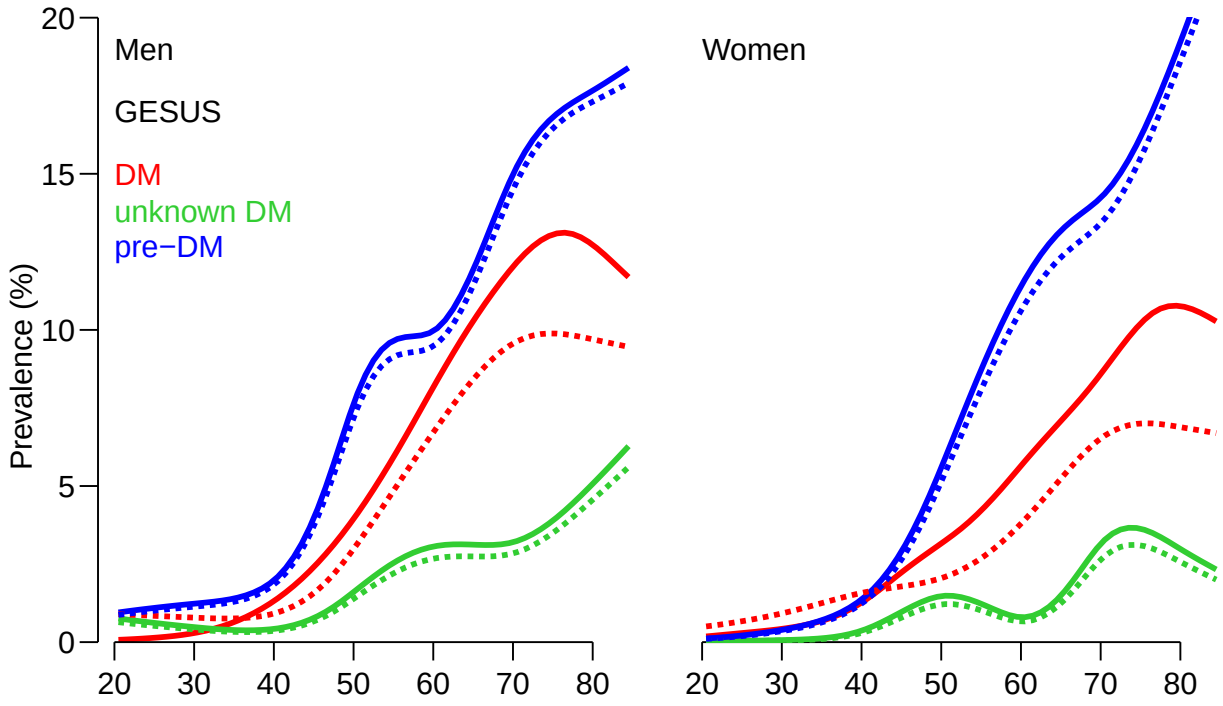

Figure 4.3: *Estimated age-specific prevalences of DM, uknown DM and pre-diabetes in men (left) and women (right) in Denmark in 2011, based on the GESUS study, using shape of responses (0,1,2,3). Median date of survey is 2012.*
Full lines are estimates of the population prevalences *and the broken lines 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`

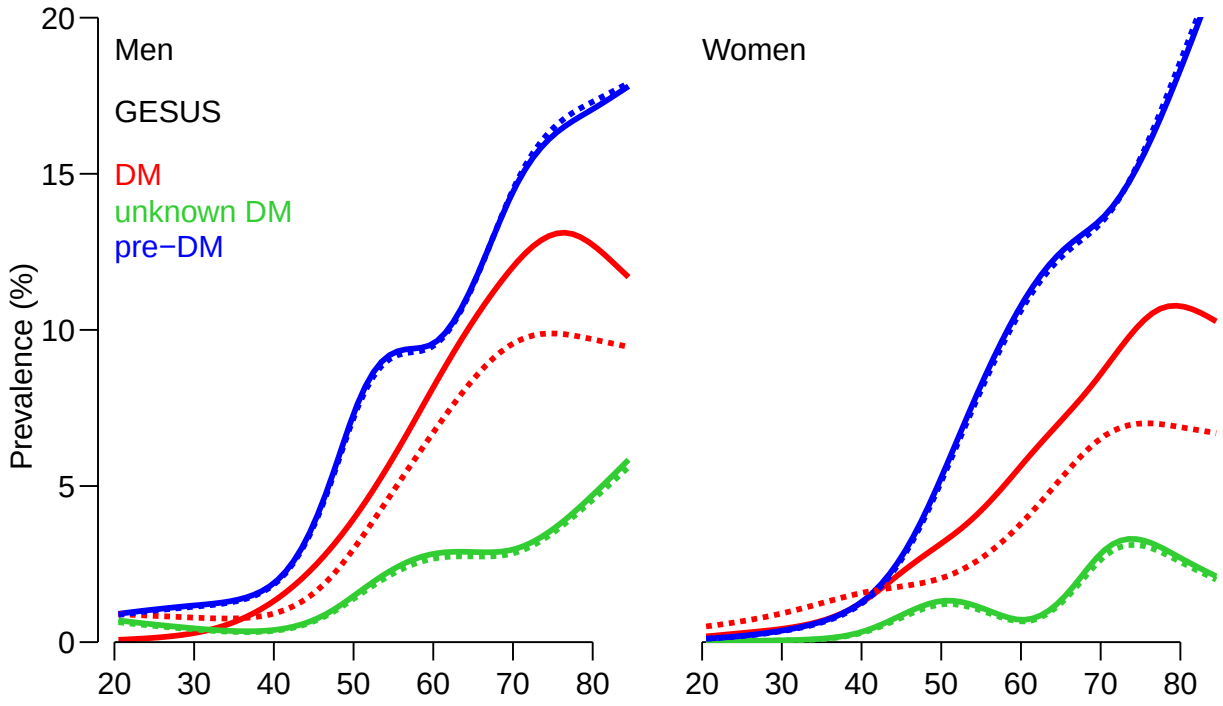

Figure 4.4: *Estimated age-specific prevalences of DM, uknown DM and pre-diabetes in men (left) and women (right) in Denmark in 2011, based on the GESUS study, using shape of responses (0,1,1.3,1.5). Median date of survey is 2011.5.*
Full lines are estimates of the population prevalences *and the broken 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`

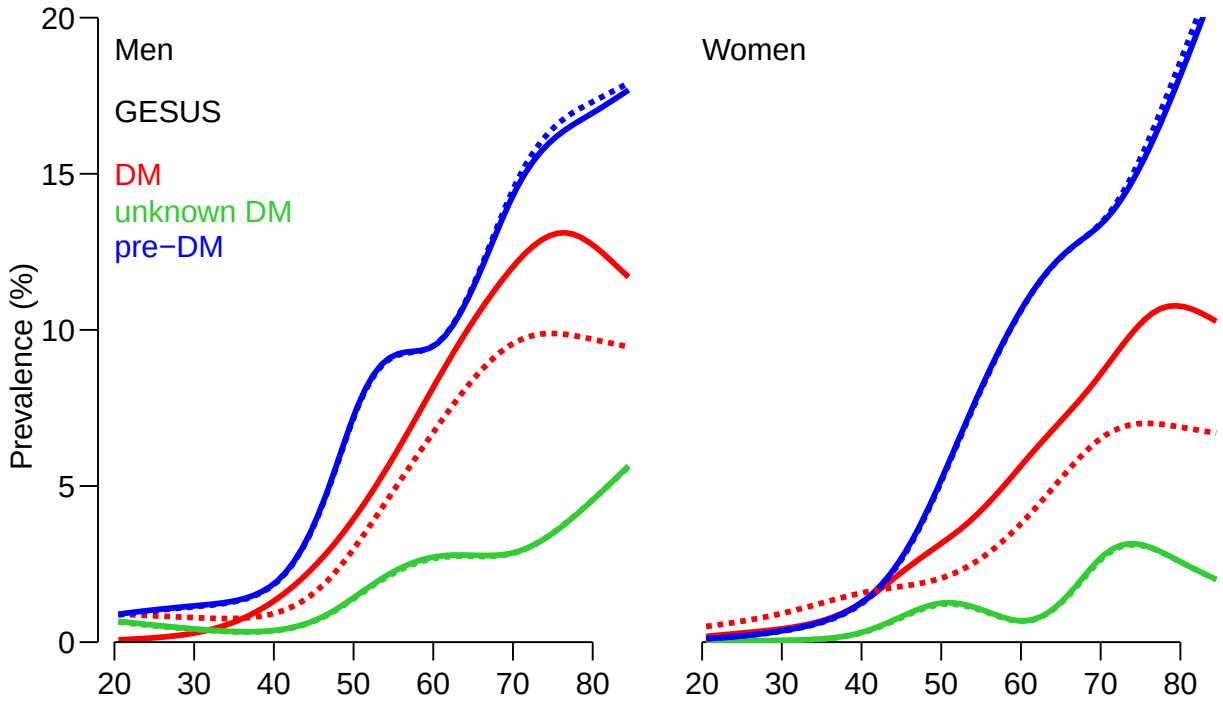

Figure 4.5: *Estimated age-specific prevalences of DM, uknown DM and pre-diabetes in men (left) and women (right) in Denmark in 2011, based on the GESUS study, using shape of responses (0,1,1.1,1.2). Median date of survey is 2011.5.*
Full lines are estimates of the population prevalences *and the broken 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")
```

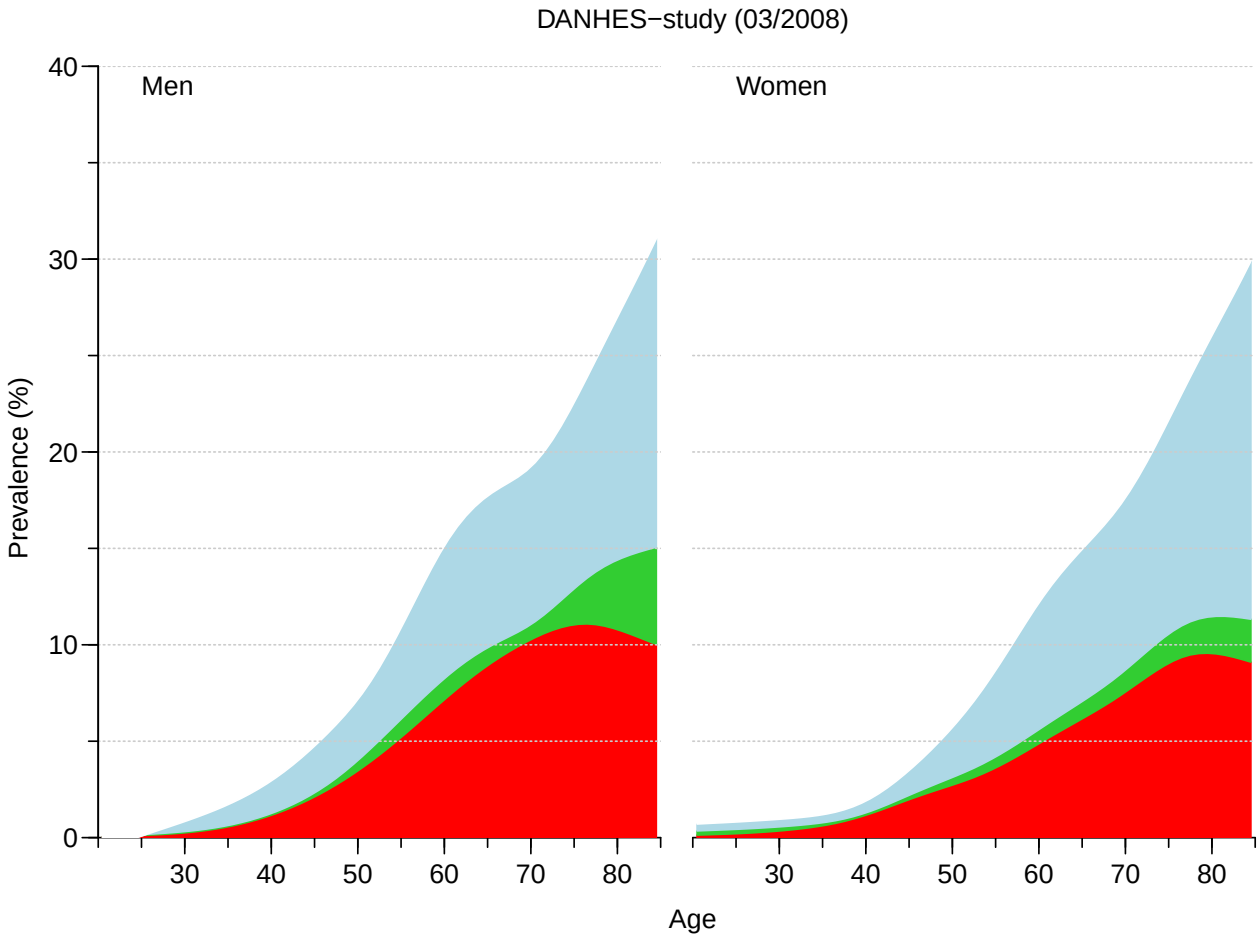

Figure 4.6: *Estimated age-specific prevalences of DM (red), uknown DM (green), pre-diabetes (light blue) and no diabetes (white) in men (left) and women (right) in Denmark in 2011, based on the DANHES study. Median date of survey is 2008.3.* `../graph/DF-stDH`

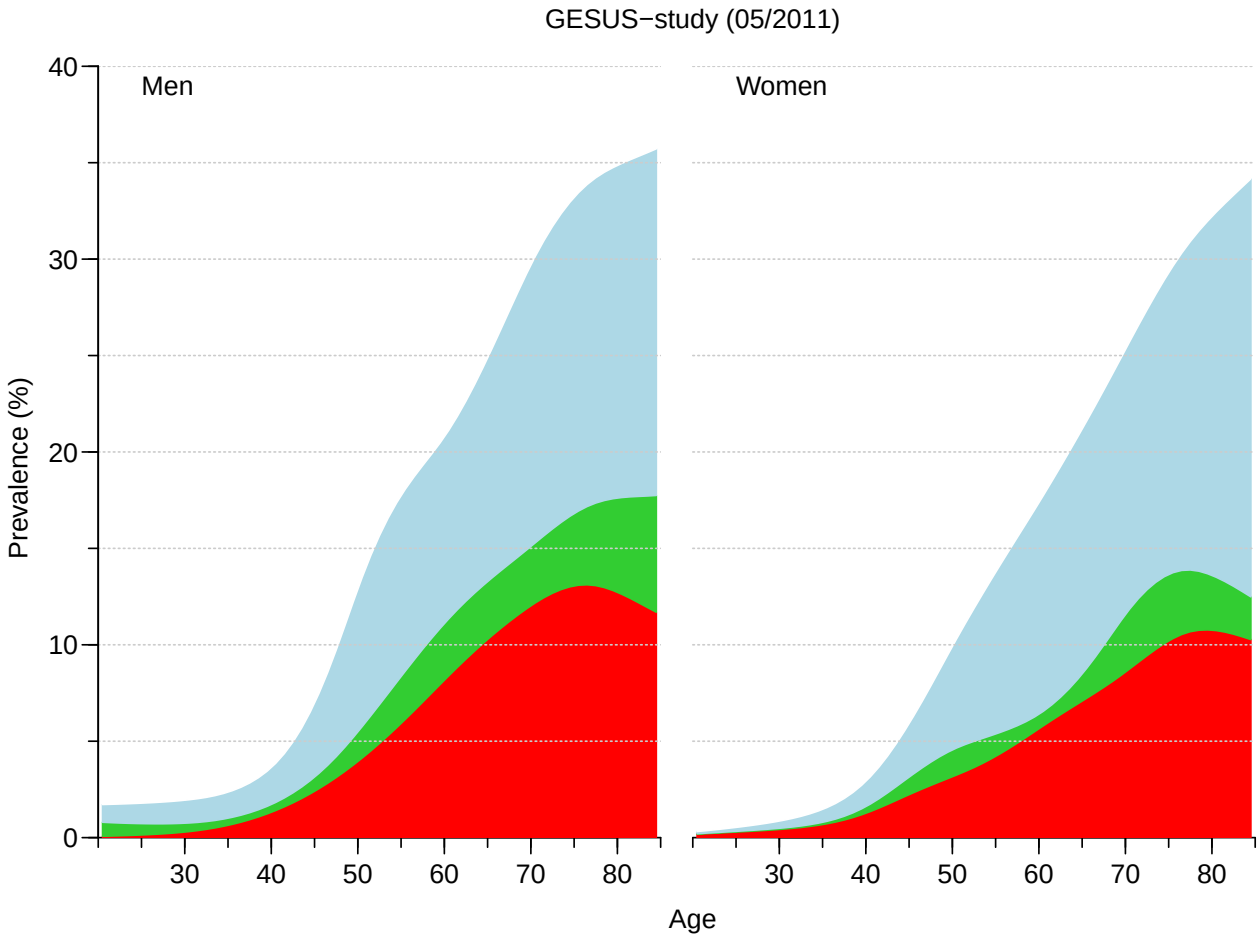

Figure 4.7: *Estimated age-specific prevalences of DM (red), uknown DM (green), pre-diabetes (light blue) and no diabetes (white) in men (left) and women (right) in Denmark in 2011, based on the GESUS study. Median date of survey is 2011.5.* `../graph/DF-stGes`

4.4 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"
```

4.4.1 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
```

5

EASD presentations

5.1 EASD abstract

5.1.1 Background and aim

Up-to-date information on unknown diabetes (DM) and prediabetes based on current diagnostic criteria as well as estimates for the entire adult age span are lacking. The aim of the study was to model the number of individuals with unknown diabetes and prediabetes in Denmark based on existing population based cohorts, correcting for differential response rates between person with and without disease.

5.1.2 Methods

Four population-based Danish studies after 2000 were identified where information on HbA1c, date of examination, gender, age (date of birth), and known diabetes was available. It is known that response rates in surveys are smaller from person with disease than for persons without, so prevalences from the surveys will underestimate the prevalence of DM and possibly also unkown DM and pre-diabetes. Within each study we estimated the age-specific prevalences of DM, unknown DM and pre-diabetes (survey prevalences), and from a National Diabetes Register we estimated the population level age-specific prevalences of known DM. We could then esitmate the survey participation rate among DM patients, and when combining this with the known overall response rates from the studies we could correct the survey prevalences to the credible population level figures.

5.1.3 Results

The prevalence of known-, previously unknown- and pre-diabetes was highest among men and increased with age with a peak at age 70 (see figure)[1]. The prevalence of undiagnosed diabetes is about half of that of diagnosed diabetes, and the prevalence of pre-diabetes is close to the prevalence of diabetes among men, but substantially higher among women.

[1] Abstract figure is Figure 4.2.

5.1.4 Conclusion

The estimated numbers with undiagnosed diabetes and prediabetes are markedly lower than suggested by previous studies, but it is not clear whether this reflects a true fall in incidence or the change to HbA1c-based diagnostic criteria in 2011. Correction for non-differential survey participation is essential in studies of diabetes and precursors.