Short Course in Epidemiology
Advanced Stream

Bendix Carstensen — Clinical Epidemiology
Senior Statistician — Steno Diabetes Center, Gentofte, Denmark
& Department of Biostatistics, University of Copenhagen
bxc@steno.dk
http://BendixCarstensen.com

Ed Gregg — Epidemiology and Statistics Branch
Chief — Division of Diabetes Translation
Centers for Disease Control, Atlanta, USA

IDEG, Vancouver, BC, Canada

December 2016

Advanced stream overview

- 0a: About the stream and practicals
- 0b: Tally of computing resources
- Ia: Population Sureveillance &cetera (EG)
- Ib: Population based Epidemiology, Prevalence (BxC)
- Ic: Prevalence practical (Section 1.2, p. 2)
- Break
- IIa: US Case examples (EG)
- IIb: Population dynamics, rates (BxC)
- IIc: Incidence practical (section 1.3, p. 4)
- IId: Mortality, survival and adherence to this world (BxC)
- III: Commercial

Population based epidemiology

Short Course in Epidemiology
Advanced Stream
December 2016
IDEG, Vancouver, BC, Canada
http://BendixCarstensen.com/Epi/Courses/IDEG2015/

epi-intro

Population based epidemiology

- What goes on in the population **as such**
- Descriptive epidemiology
- Medical demography:
  - how many are affected by diabetes
    — at a given date
  - how many acquire diabetes
    — in a given timespan
  - how many die (w/o diabetes)
    — in a given timespan
- Prevalence
- Incidence
- Mortality

Types of measures

- Status measures:
  what is the **state of affairs** at a given date (or age, say)
  Prevalence
- Flow measures:
  what are **rates of change** at a given date (or age, say)
  Incidence, Mortality

Prevalence

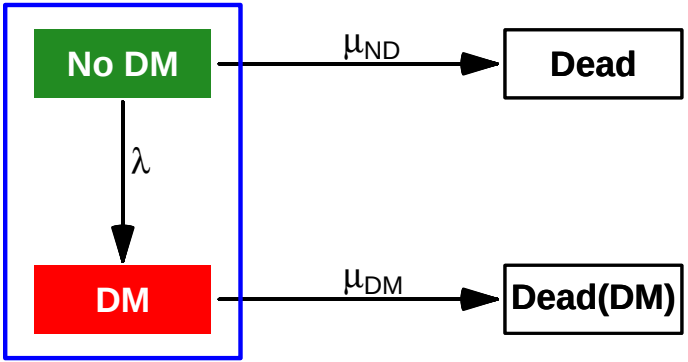
Short Course in Epidemiology
Advanced Stream
December 2016
IDEG, Vancouver, BC, Canada
http://BendixCarstensen.com/Epi/Courses/IDEG2015/

prev-ana

Prevalence

- What is the fraction with diabetes among all alive at a given date
  — **empirical** prevalence
- What is the probability that a randomly chosen (living) person has diabetes
  — **theoretical** prevalence

Components of diabetes prevalence



Prevalence — data

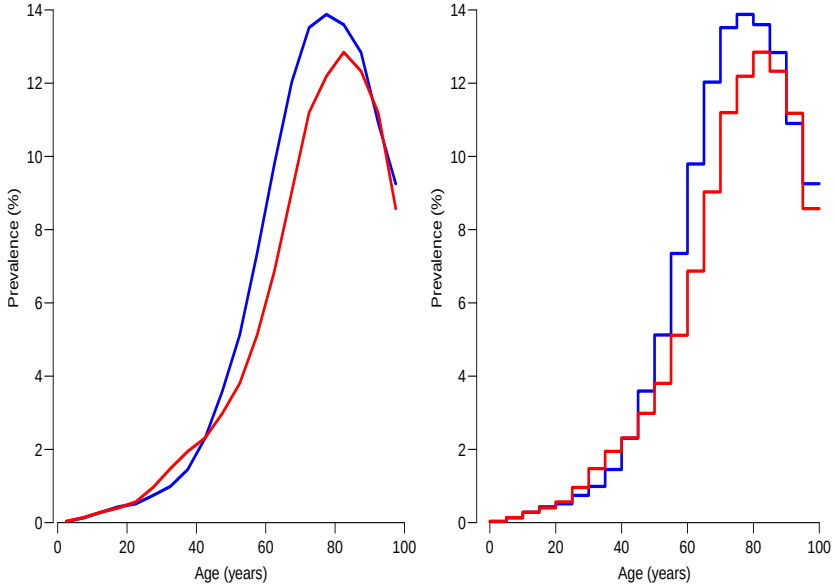
- **empirical** prevalence:
  - ratio of counts (affected / total)
  - requires a specific population
    (*e.g.* 50–59 year old women at 1 Jan 2008)
- **theoretical** prevalence
  - requires a statistical model for binomial outcome, *e.g.*

$$\mathrm{P}\{\mathrm{DM}\} = f(a, p)$$

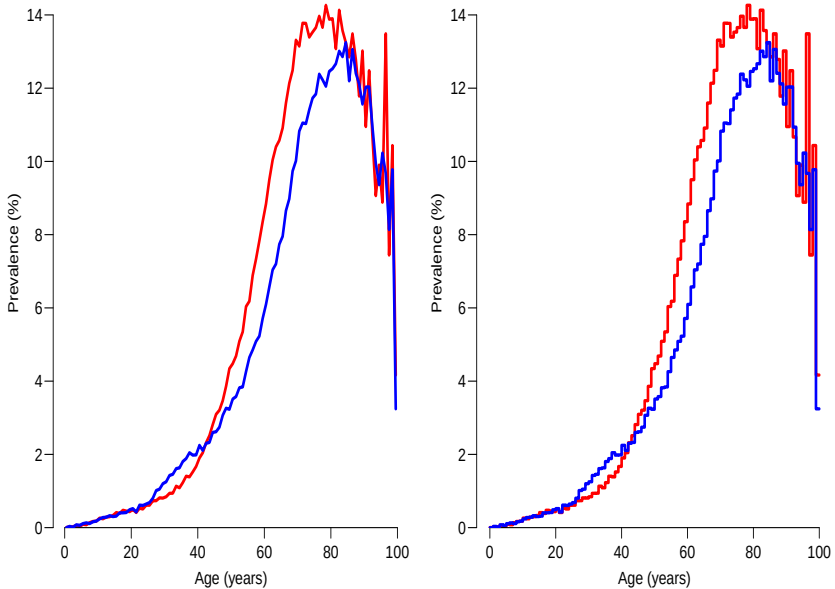
  - "this is how data **could** have been generated"
  - **Estimated** probability...
  - based on data (X, N) for different values of a and p

Your turn — 1.2 Prevalence, p. 2

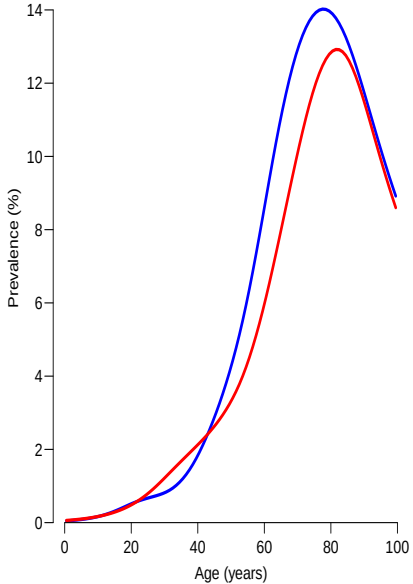
Analyzing prevalence

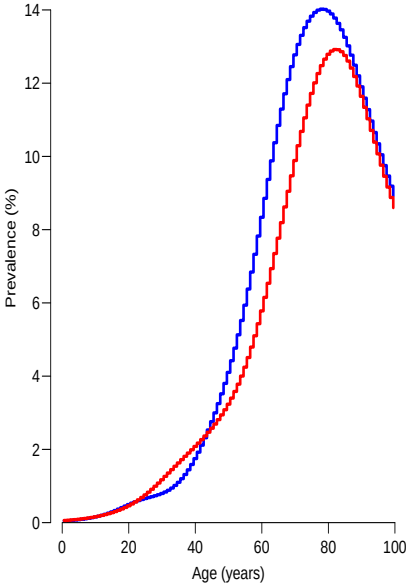


Analyzing prevalence



Analyzing prevalence





Analyzing prevalence

Assumptions in statistical model:

- Prevalence varies **continuously** with age
- Prevalence varies **smoothly** with age

Data for a statistical model:

- No. prevalent DM cases in age classes
- No. total population in age classes
- Quantitative age in each age class (midpoint, *e.g.*)

— age classes as small as possible
— modeling smooths the random variation between classes.

Population dynamics

Short Course in Epidemiology
Advanced Stream
December 2016
IDEG, Vancouver, BC, Canada
http://BendixCarstensen.com/Epi/Courses/IDEG2015/

pop

What goes on in a population?

- Birth, death
- Birth, disease, death
- Birth, disease (duration), death
- Diseased persons (prevalence (p)):
  - increases by new cases (incidence rate, λ)
  - decreases by cases dying (mortality rate, μ)
  - increases by duration from disease to death, y
    — the longer duration the higher the prevalence.
- In a small interval $\mathrm{d}t$:

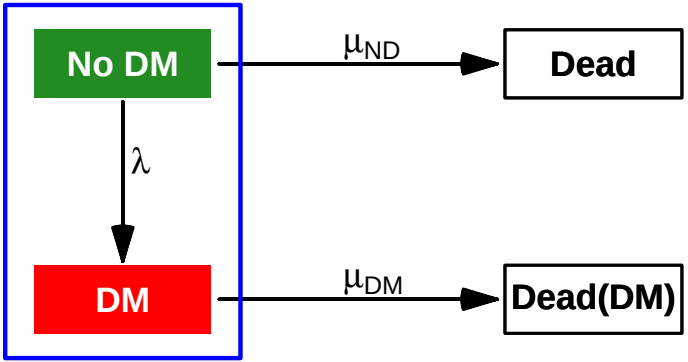
$$p(t + \mathrm{d}t) = p(t) - \mu\,\mathrm{d}t + \lambda\,\mathrm{d}t$$

- No change in prevalence if $\lambda = \mu$

Relationships

- $p = \lambda \times y$:
  prevalence equals incidence times duration.
  - X new cases per population N per year
  - $\lambda = X/N$
  - each case lasting y years: Xy cases present
  - Prevalence $p = Xy/N = \lambda y$
- — only in steady state
- — only if independent of age
- Mainly a **qualitative** relationship
- because populations are not in steady state.

Components of diabetes prevalence



Quantities determining prevalence are the rates

Rates determine all

If we know the rates, then we know:

- prevalence
- survival
- lifespan with DM
- lifespan without DM
- years of life lost
- lifespan with DM and CVD
- etc. . .

. . . meaning that we can **simulate** everything from rates.

Note: rates depend on age, duration, calendar time . . .

Incidence rates

Short Course in Epidemiology
Advanced Stream
December 2016
IDEG, Vancouver, BC, Canada
http://BendixCarstensen.com/Epi/Courses/IDEG2015/

inc-ana

Incidence rate of DM

- **Rate**: events per time
- **Events**: New cases of DM
- **(risk) Time**: Time lived by persons susceptible to the event

The DM register

One line is one person:

```
> load( file="../data/dr.Rda" )
> nrow( subset( dr, floor(doDM)==2006 &
+                   (doDM-doBth)>=60 &
+                   (doDM-doBth)< 65 ) )
[1] 3480
```

Population data I

One line is one age×period class

```
> load( file="../data/Ydk.Rda" )
> str( Ydk )
'data.frame': 8400 obs. of  4 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  1971 1971 1971 1971 1971 ...
 $ Y  : num  37139 35129 36134 34223 37113 ...
> head( Ydk )
  sex A    P        Y
1   M 0 1971 37139.17
2   F 0 1971 35128.83
3   M 1 1971 36133.67
4   F 1 1971 34223.00
5   M 2 1971 37113.00
6   F 2 1971 34926.33
```

Population data II

```
> subset( Ydk, A>=60 & A<65 & P==2006 )
     sex  A    P        Y
7121   M 60 2006 40160.17
7122   F 60 2006 39678.17
7123   M 61 2006 38069.33
7124   F 61 2006 37952.83
7125   M 62 2006 35100.50
7126   F 62 2006 35562.67
7127   M 63 2006 32311.67
7128   F 63 2006 32980.00
7129   M 64 2006 29321.67
7130   F 64 2006 30069.83
> sum( subset( Ydk, A>59 & A<65 & P==2006 )$Y )
[1] 351206.8
```

Events by age and sex I

Tables of events and PY by sex:

```
> ( D <- with( subset(dr,floor(doDM)==2006),
+              table( floor((doDM-doBth)/5)*5, sex ) ) )
```

```
     sex
        M    F
  0    26   29
  5    41   38
  10   79   85
  15   55  101
  20   73  101
  25  107  158
  30  178  275
  35  352  317
  40  597  608
  45  882  614
  50 1307  894
  55 1659 1152
```

Events by age and sex II

60	2038	1442
65	1583	1342
70	1195	1175
75	950	1096
80	634	834
85	245	470
90	70	153
95	12	31
100	0	2

Person-years by age and sex I

Tables of events and PY by sex:

```
> ( Y <- xtabs( Y ~ I(floor(A/5)*5) + sex, data=subset(Ydk,P==2006) ) )
                  sex
I(floor(A/5) * 5)          M          F
               0  166333.333 158679.833
               5  173061.833 164956.500
               10 180623.833 171379.833
               15 163695.000 154952.667
               20 149068.000 144681.333
               25 164809.667 163738.500
               30 191372.167 189887.167
               35 200951.833 194950.833
               40 212268.000 205365.000
               45 188801.833 184550.167
               50 181232.333 179168.000
               55 186422.833 186106.500
               60 174963.333 176243.500
```

Person-years by age and sex II

```
65 121788.167 129678.000
70  91038.500 105248.833
75  68313.833  88990.167
80  45502.167  73541.667
85  22305.833  47119.833
90   7295.500  20757.833
95   1413.667   6093.667
```

Incidence data I

Divide tables of events and PY by sex:

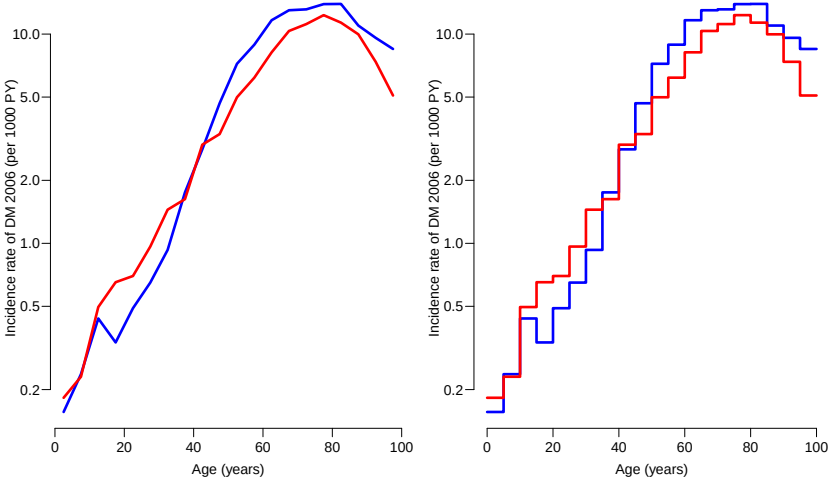
```
> D <- D[1:20,]
> inc <- D/Y * 1000
> round( cbind( D, Y, inc ), 3 )
```

	M	F	M	F	M	F
0	26	29	166333.333	158679.833	0.156	0.183
5	41	38	173061.833	164956.500	0.237	0.230
10	79	85	180623.833	171379.833	0.437	0.496
15	55	101	163695.000	154952.667	0.336	0.652
20	73	101	149068.000	144681.333	0.490	0.698
25	107	158	164809.667	163738.500	0.649	0.965
30	178	275	191372.167	189887.167	0.930	1.448
35	352	317	200951.833	194950.833	1.752	1.626
40	597	608	212268.000	205365.000	2.812	2.961
45	882	614	188801.833	184550.167	4.672	3.327
50	1307	894	181232.333	179168.000	7.212	4.990
55	1659	1152	186422.833	186106.500	8.899	6.190

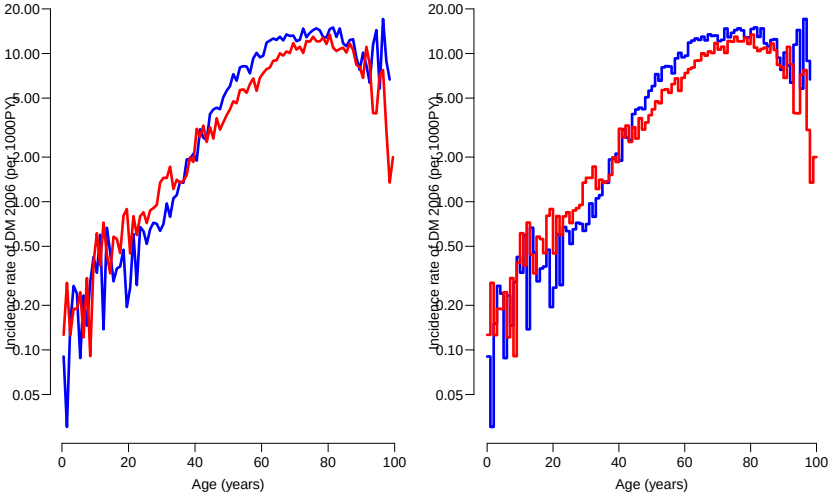
Incidence data II

60	2038	1442	174963.333	176243.500	11.648	8.182
65	1583	1342	121788.167	129678.000	12.998	10.349
70	1195	1175	91038.500	105248.833	13.126	11.164
75	950	1096	68313.833	88990.167	13.906	12.316
80	634	834	45502.167	73541.667	13.933	11.341
85	245	470	22305.833	47119.833	10.984	9.975
90	70	153	7295.500	20757.833	9.595	7.371
95	12	31	1413.667	6093.667	8.489	5.087

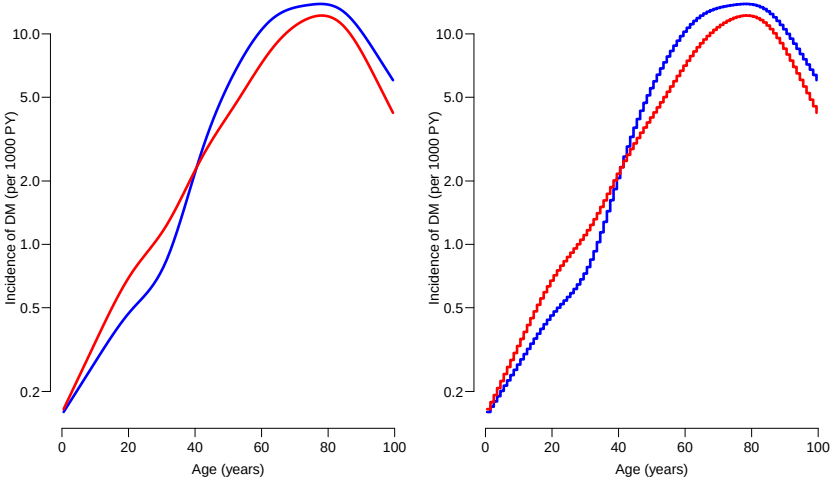
Incidence curves



Incidence curves



Incidence curves



People only get diabetes once... I

- In our calculations we used the **entire** population-PY as denominator
- ...but it should only be those **susceptible** to DM,
- that is, those without a diagnosis.
- So we should subtract the PY among in DM-persons from the population PY

People only get diabetes once. . . I

```
> dmY <- Y * 0
> for( sx in c("M","F") )
+ for( aa in 1:20 )
+ dmY[aa,sx] <- with( subset(dr,sex==sx),
+                     sum( pmax( 0, pmin(2007,doBth+ aa   *5,doDth,na.rm=TRUE) -
+                                   pmax(2006,doBth+(aa-1)*5,doDM ) ) ) )
> round( 100*t(dmY/Y), 1 )
   I(floor(A/5) * 5)
sex    0    5   10   15   20   25   30   35   40   45   50   55   60   65   70   7
  M  0.0  0.1  0.3  0.4  0.5  0.7  1.0  1.5  2.3  3.8  5.5  7.7 10.6 12.8 14.9 15.
  F  0.0  0.1  0.3  0.4  0.6  0.9  1.5  2.1  2.5  3.2  4.2  5.5  7.3  9.6 12.1 13.
   I(floor(A/5) * 5)
sex   95
  M 9.3
  F 8.3
```

Incidence data I

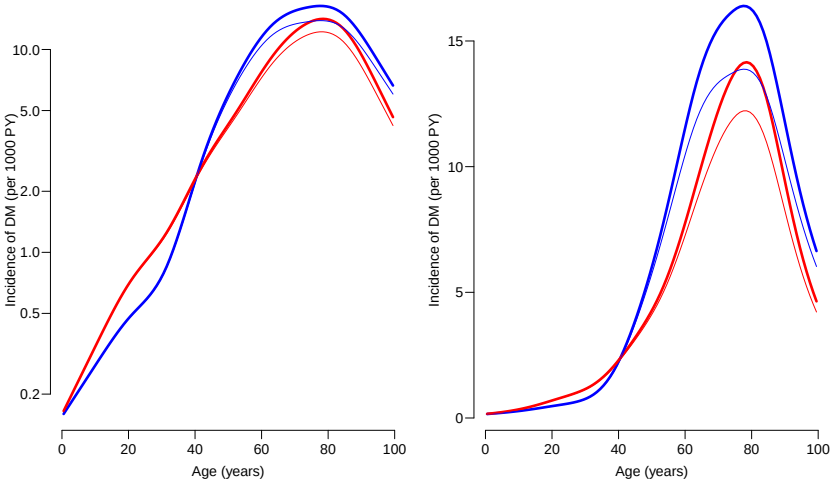
Divide tables of events and PY by sex:

```
> round( cbind( D, Y-dmY, inc=D/(Y-dmY)*1000 ), 3 )
      M    F          M          F      M      F
0    26   29 166271.972 158616.314  0.156  0.183
5    41   38 172835.243 164710.285  0.237  0.231
10   79   85 180075.719 170846.886  0.439  0.498
15   55  101 162961.780 154280.319  0.338  0.655
20   73  101 148291.398 143828.127  0.492  0.702
25  107  158 163576.325 162209.046  0.654  0.974
30  178  275 189413.859 186995.975  0.940  1.471
35  352  317 197867.377 190940.455  1.779  1.660
40  597  608 207282.839 200196.856  2.880  3.037
45  882  614 181720.264 178650.303  4.854  3.437
50 1307  894 171227.261 171672.990  7.633  5.208
55 1659 1152 172004.168 175814.414  9.645  6.552
60 2038 1442 156393.482 163348.876 13.031  8.828
65 1583 1342 106147.963 117196.741 14.913 11.451
```

Incidence data II

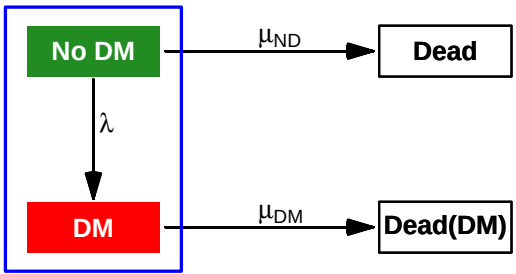
70	1195	1175	77472.789	92552.577	15.425	12.695
75	950	1096	58019.610	77074.413	16.374	14.220
80	634	834	38714.886	63474.716	16.376	13.139
85	245	470	19202.411	40731.951	12.759	11.539
90	70	153	6412.488	18248.537	10.916	8.384
95	12	31	1281.738	5585.846	9.362	5.550

Incidence curves



Analysis of rates

- Get the denominator right
- Count only the transitions of interest



Your turn — 1.3 Incidence, p 4.

Mortality rates

Short Course in Epidemiology
Advanced Stream
December 2016
IDEG, Vancouver, BC, Canada
http://BendixCarstensen.com/Epi/Courses/IDEG2015/

mort-ana

Mortality

- Mortality rates $\leftrightarrow$ Survival
- . . . but **only** if we have an origin
- **rates** are instantaneous — independent of previous history
- **survival** is a cumulative measure:
- time **since** some time:
  - birth
  - DM diagnosis
  - Start of treatment
  - Inclusion in clinical trial
  - Inclusion in cohort study — meaningful?

```
> library( Epi )
> load( file="../data/dr.Rda" )
> load( file="../data/dmY.Rda" )
> dd <- with( subset( dr, floor(doDth)==2006 ),
+             table( floor(doDth-doBth), sex ) )
> str( dd )
 'table' int [1:88, 1:2] 0 1 1 0 2 0 1 1 0 2 ...
 - attr(*, "dimnames")=List of 2
  ..$    : chr [1:88] "0" "12" "18" "20" ...
  ..$ sex: chr [1:2] "M" "F"
> dmD <- dmY * 0 # a table of 0s with same structure as the person-years
> for( aa in intersect( dimnames(dd )[[1]], # fill in deaths where they are
+                       dimnames(dmD)[[1]] ) )
+      dmD[aa,] <- dd[aa,]
> cbind( dmD, dmY )
```

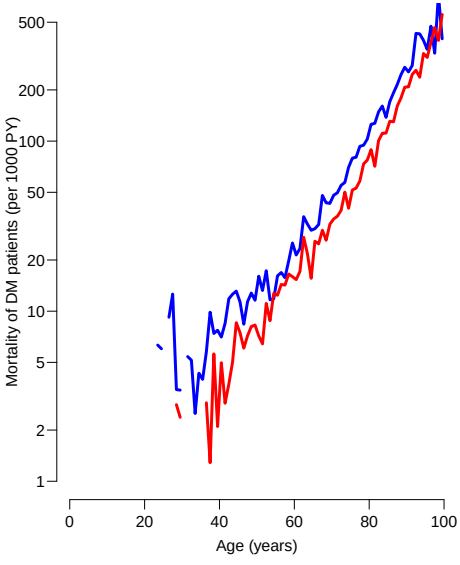
	M	F	M	F
0	0	2	0.5991139	0.8834697
1	0	0	2.2238462	5.1816978
2	0	0	10.0438410	14.6850286
3	0	0	17.0520062	16.4032660
4	0	0	31.4427208	26.3662713
5	0	0	25.4418535	36.5189540
6	0	0	34.6849836	40.7997248
7	0	0	41.2744337	51.0909042
8	0	0	52.6381848	56.9708197
9	0	0	72.5509094	60.8343396
10	0	0	85.3786136	85.2971495
11	0	0	97.2820296	92.0268266
12	1	0	120.0022359	111.8253605
13	0	0	115.8396328	129.0822457
14	0	0	129.6119440	114.7156362
15	0	0	131.8663390	121.9779585
16	0	0	137.7102903	121.6464603
17	0	0	146.4369793	137.8939167
18	1	0	157.8277979	142.3506072
19	0	0	159.3784021	148.4783878
20	0	1	149.7101520	154.7186798

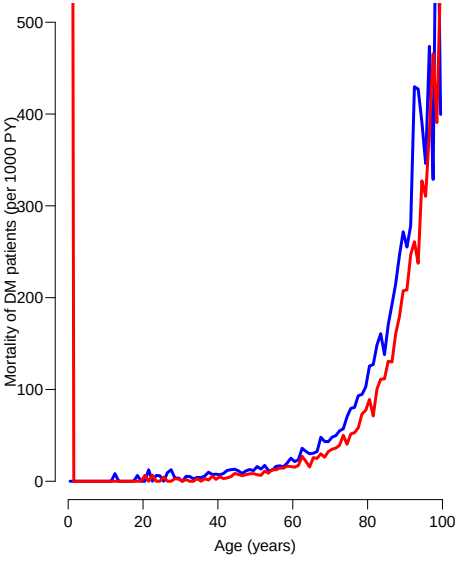
21	2	0	159.8240811	167.7973905
22	0	1	142.4998350	147.8355276
23	1	0	158.2059093	181.0933141
24	1	0	166.3616880	201.7611598
25	0	1	198.0180497	219.7468375
26	2	0	217.0691592	241.7829944
27	3	0	238.0459089	292.6649769
28	1	1	289.0175594	354.5225200
29	1	1	291.1906626	420.7362345
30	0	0	337.3480178	465.0634935
31	2	1	369.8610480	535.5297856
32	2	0	388.2872535	575.6638541
33	1	0	399.5219924	644.9040345
34	2	2	463.2897195	670.0309441
35	2	0	503.4987334	697.1798589
36	3	2	519.7317942	691.0730864
37	6	1	608.5930030	777.7685538
38	5	5	675.9224563	891.2704106
39	6	2	776.7104921	953.0860758
40	6	5	851.7037109	1005.1466880
41	8	3	943.1357642	1043.2090555
42	12	4	1014.2465616	1074.4989623

```
43  13   5 1032.9122856  993.8265132
44  15   9 1143.1631323 1051.4625842
45  14   8 1236.7600363 1071.9565711
46  11   7 1311.7568816 1154.2537850
47  16   8 1408.1272657 1116.3735545
48  19  10 1486.4897341 1230.9328330
49  19  11 1638.4353889 1326.3465553
50  29  10 1806.6080278 1398.7197564
51  25   9 1889.1044999 1401.0834764
52  34  17 1965.9349339 1530.4011980
53  25  14 2144.9687562 1591.7855328
54  26  20 2198.4565389 1573.0196803
55  38  21 2356.6929735 1692.9809287
56  43  27 2547.1730908 1881.0030712
57  44  29 2794.6999525 2036.3213516
58  63  37 3185.4952440 2237.9205234
59  89  39 3534.6040233 2443.8599857
60  82  39 3830.5961627 2543.1885368
61  90  45 3853.8911425 2623.7722501
62 136  71 3782.6235474 2608.7045714
63 119  57 3657.8934755 2619.3721537
64 103  39 3444.8469698 2499.5860958
```

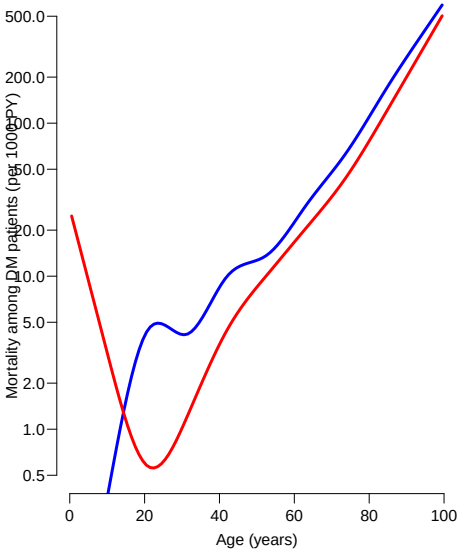
```
65  97  63 3175.0308881 2444.3942450
66 104  60 3218.0856812 2411.1838445
67 147  74 3067.7435265 2471.9917799
68 136  67 3135.4591450 2565.0095692
69 131  84 3043.8849018 2588.6799548
70 141  90 2935.6734375 2583.8702941
71 145  93 2919.1940537 2572.7331335
72 150 101 2729.4571012 2578.1553072
73 147 123 2572.6509370 2454.7790567
74 169 101 2408.7354579 2506.7189450
75 180 129 2268.5285174 2500.4774297
76 175 127 2177.3356298 2400.9440466
77 197 141 2113.8368364 2421.5974775
78 183 172 1935.0518805 2343.8443926
79 185 174 1799.4706737 2248.8899836
80 212 196 1689.5302632 2199.1535400
81 195 154 1533.6681125 2165.9042405
82 197 203 1326.5886888 2017.5379713
83 186 209 1157.1877826 1880.7084905
84 149 201 1080.3061885 1803.6466494
85 161 227  944.8050421 1736.7338926
86 145 195  753.8669682 1501.4529501
```

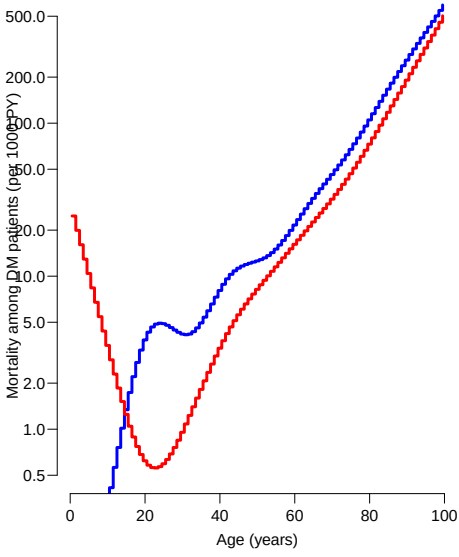
```
87 123 195  571.6494811 1213.5076211
88 118 190  479.8750475 1059.2705835
89  96 182  353.2254999  876.9174655
90  76 148  297.9148159  710.0238486
91  63 152  226.3693954  616.3634914
92  72 136  167.5491735  521.1093040
93  50  94  117.0354553  395.9485944
94  29  87   74.1429975  265.8511126
95  19  58   54.8777396  186.8722191
96  17  51   35.8853082  134.1148951
97   7  44   21.2913490   94.5064904
98   9  22   12.3662040   56.3150190
99   3  20    7.5076243   36.0117959
```





. . . **not** credible


500.0
200.0
100.0
50.0
20.0
10.0
5.0
2.0
1.0
0.5
Mortality among DM patients (per 1000PY)
0
20
40
60
80
100
Age (years)


500.0
200.0
100.0
50.0
20.0
10.0
5.0
2.0
1.0
0.5
Mortality among DM patients (per 1000PY)
0
20
40
60
80
100
Age (years)

Mortality and duration I

Not meaningful without conditioning on age

What is the survival of a person diagnosed at age 60?

```
> dr60 <- subset( dr, floor(doDM-doBth)== 60 &
+                           doDM       >1995 )
> D60 <- with( dr60, table( floor(doDth-doDM), sex ) )
```

This is deaths among persons diagnosed in age 60

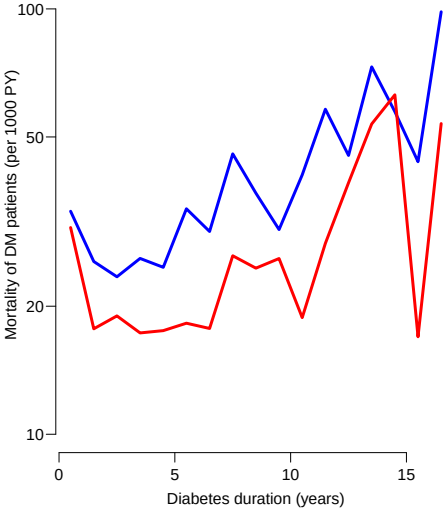
Mortality and duration I

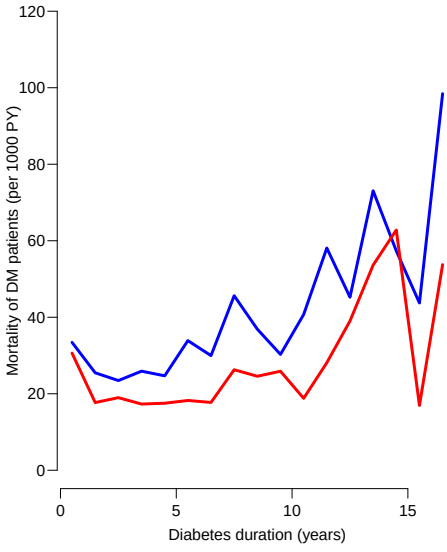
But we also need PY

```
> Y60 <- D60 * 0
> for( sx in dimnames(Y60)[[2]] )
+ for( dd in dimnames(Y60)[[1]] )
+ Y60[dd,sx] <- with( subset( dr60, sex==sx ),
+                     sum( pmax( pmin(2012,   # end of FU
+                                     doDth,  # in duration dd
+                                     doDM+as.numeric(dd)+1,
+                                     na.rm=TRUE) -
+                                (doDM+as.numeric(dd)), # start of FU
+                                0 ) ) ) # discard negative FU
> cbind( D60, round(Y60,1), round(1000*D60/Y60,1) )
```

Mortality and duration II

	M	F	M	F	M	F
0	178	117	5323.4	3821.6	33.4	30.6
1	121	61	4749.5	3446.6	25.5	17.7
2	99	59	4220.9	3109.0	23.5	19.0
3	97	48	3744.7	2773.2	25.9	17.3
4	81	43	3280.8	2453.7	24.7	17.5
5	96	39	2832.6	2137.8	33.9	18.2
6	72	33	2402.2	1861.3	30.0	17.7
7	89	41	1950.7	1560.2	45.6	26.3
8	58	31	1572.6	1261.4	36.9	24.6
9	37	26	1221.6	1004.3	30.3	25.9
10	39	15	957.8	797.7	40.7	18.8
11	43	18	740.2	640.4	58.1	28.1
12	25	19	552.6	487.0	45.2	39.0
13	28	18	383.2	335.6	73.1	53.6
14	14	13	244.1	207.0	57.3	62.8
15	6	2	137.2	118.0	43.7	17.0
16	4	2	40.6	37.2	98.4	53.7


100
50
20
10
Mortality of DM patients (per 1000 PY)
0
5
10
15
Diabetes duration (years)


120
100
80
60
40
20
0
Mortality of DM patients (per 1000 PY)
0
5
10
15
Diabetes duration (years)

Survival curve I

$$S(t) = (1 - \mu_0) \times (1 - \mu_1) \times (1 - \mu_2) \times \cdots$$

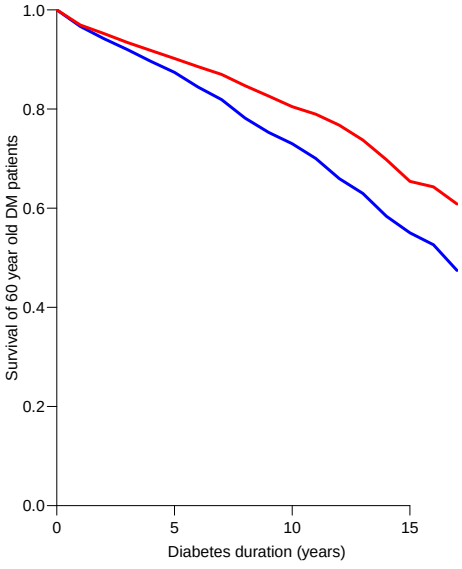
```
> ( p60 <- 1 - D60/Y60 )
    sex
           M         F
0  0.9665625 0.9693842
1  0.9745237 0.9823014
2  0.9765453 0.9810227
3  0.9740970 0.9826913
4  0.9753110 0.9824754
5  0.9661094 0.9817572
6  0.9700272 0.9822702
7  0.9543746 0.9737208
8  0.9631180 0.9754235
```

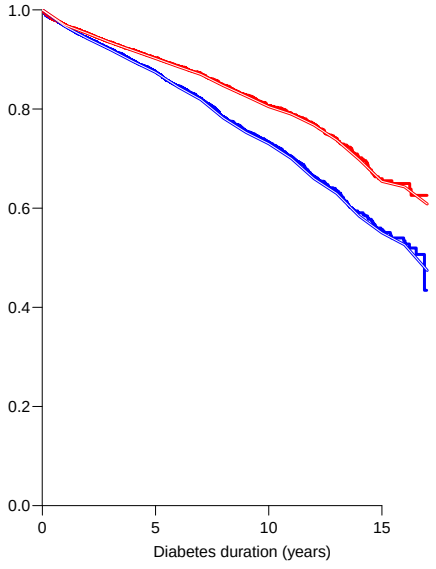
Survival curve II

```
   9  0.9697122 0.9741110
   10 0.9592825 0.9811961
   11 0.9419072 0.9718911
   12 0.9547620 0.9609833
   13 0.9269371 0.9463654
   14 0.9426528 0.9371992
   15 0.9562704 0.9830491
   16 0.9015634 0.9462515
> ( S60 <- rbind( 1, apply( p60, 2, cumprod ) ) )
            M         F
    1.0000000 1.0000000
0   0.9665625 0.9693842
1   0.9419381 0.9522275
2   0.9198452 0.9341568
3   0.8960185 0.9179877
4   0.8738967 0.9019004
5   0.8442798 0.8854472
6   0.8189743 0.8697484
7   0.7816083 0.8468921
```

Survival curve III

```
 8  0.7527811 0.8260784
 9  0.7299810 0.8046921
10 0.7002580 0.7895607
11 0.6595780 0.7673671
12 0.6297401 0.7374269
13 0.5837294 0.6978753
14 0.5502542 0.6540482
15 0.5261918 0.6429616
16 0.4743952 0.6084034
```


1.0
0.8
0.6
0.4
0.2
0.0
Survival of 60 year old DM patients
0
5
10
15
Diabetes duration (years)


1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
Diabetes duration (years)

Mortality and survival

- Formally there is a one-to-one correspondence
- But survival requires a meaningful **origin**
- . . . such as date of diagnosis or date of start of intervention
- Survival is time **since**. . .
- . . . time since inclusion in a cohort is not meaningful

Population measures

Short Course in Epidemiology
Advanced Stream
December 2016
IDEG, Vancouver, BC, Canada
http://BendixCarstensen.com/Epi/Courses/IDEG2015/

measure-links

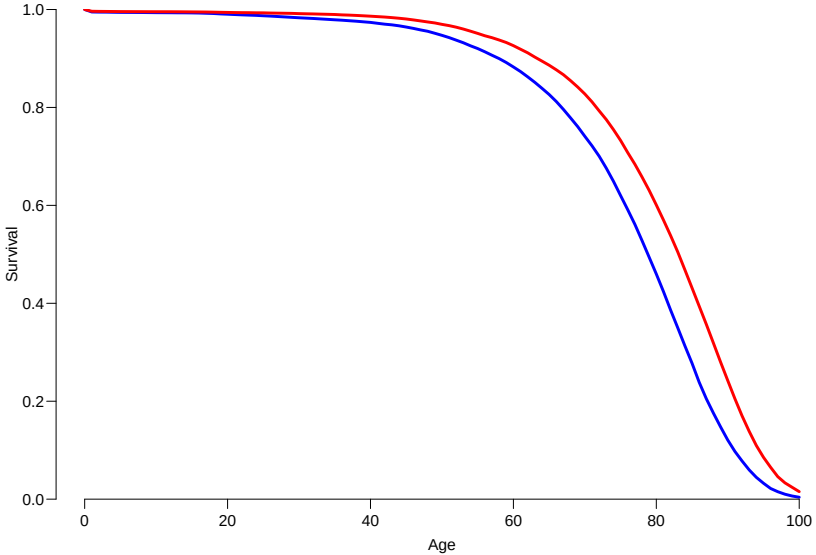
Demographic measures

- Practicals chapter 2: “Basic concepts. . . ”
- Rates
- Survival
- Competing risks
- Lifetime risk of . . .
- Years lost to . . .
- Time spent with. . .

Survival

- Normally estimated by the Kaplan-Meier estimator
- — or life table (actuarial) estimator
- In demography we want to address the survival of the **entire population**
- — or the population of diabetes patients
- . . . is the latter meaningful?

How does the survival function look?



Survival function in your computer I

$$S(t) = \exp\left(-\int_0^t \lambda(s)\,\mathrm{d}s\right) = \exp\big(-\Lambda(s)\big)$$

```
> library(Epi)
> data(M.dk)
> names(M.dk)
[1] "A"    "sex"  "P"    "D"    "Y"    "rate"
> head( subset(M.dk,P==2005 & sex==1) )
```

Survival function in your computer II

```
     A sex    P   D        Y       rate
63   0   1 2005 167 33047.17 5.05338330
141  1   1 2005   9 33280.33 0.27042998
219  2   1 2005   3 33253.67 0.09021562
297  3   1 2005   9 33396.00 0.26949335
375  4   1 2005   7 34101.33 0.20527057
453  5   1 2005   3 34416.33 0.08716791
```

```
> mort.m <- subset(M.dk,P==2005 & sex==1)$rate
> mort.f <- subset(M.dk,P==2005 & sex==2)$rate
> age    <- subset(M.dk,P==2005 & sex==1)$A
> Lamb.m <- cumsum(mort.m/1000)
> Lamb.f <- cumsum(mort.f/1000)
> Surv.m <- exp(-Lamb.m)
> Surv.f <- exp(-Lamb.f)
> cbind( age, mort.m, Lamb.m, Surv.m,
+             mort.f, Lamb.f, Surv.f )[1:10,]
```

Survival function in your computer III

	age	mort.m	Lamb.m	Surv.m	mort.f	Lamb.f	Surv.f
[1,]	0	5.05338330	0.005053383	0.9949594	3.57265182	0.003572652	0.9964337
[2,]	1	0.27042998	0.005323813	0.9946903	0.25196586	0.003824618	0.9961827
[3,]	2	0.09021562	0.005414029	0.9946006	0.12648288	0.003951101	0.9960567
[4,]	3	0.26949335	0.005683522	0.9943326	0.15702613	0.004108127	0.9959003
[5,]	4	0.20527057	0.005888793	0.9941285	0.15306591	0.004261193	0.9957479
[6,]	5	0.08716791	0.005975961	0.9940419	0.06069895	0.004321892	0.9956874
[7,]	6	0.08731718	0.006063278	0.9939551	0.03050765	0.004352399	0.9956571
[8,]	7	0.08596276	0.006149241	0.9938696	0.06039650	0.004412796	0.9955969
[9,]	8	0.16953702	0.006318778	0.9937011	0.08930210	0.004502098	0.9955080
[10,]	9	0.16586881	0.006484646	0.9935363	0.02917734	0.004531275	0.9954790

```
> plot( c(0,age+1), c(1,Surv.m),
+       type="l", lty=1, lwd=3, col="blue",
+       ylim=0:1, yaxs="i", xlab="Age", ylab="Survival" )
> lines( c(0,age+1), c(1,Surv.f),
+        type="l", lty=1, lwd=3, col="red" )
```

Survival function in your computer

Note:

- the survival curve is just at set of points $(a, S(a))$ where $a = 0, 1, 2, \ldots$
- the cumulative rate $\Lambda(a)$ was just computed as the cumulative sum of the mortality rates (in units of rate per 1 year):

```
> Lamb.m <- cumsum( mort.m/1000 )
```

Expected lifetime

Computed as the area under the survival curve:

$$EL = \int_0^\infty S(t)\,\mathrm{d}t$$

Expected **residual** lifetime at age a
computed as the area under the **conditional** survival curve:

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

Integrals are just sums (`cumsum()` in R)

Expected lifetime I

This is just the integral of the survival function:

```
> EL.m <- sum( Surv.m * 1 ) + 0.5
> EL.f <- sum( Surv.f * 1 ) + 0.5
> round( c(EL.m,EL.f), 2 )
[1] 75.93 80.43
```

For the expected residual lifetime at age 50 we need the **conditional** survival from age 50:

```
> cbind( age, Surv.m, Surv.f )[48:52,]
     age    Surv.m    Surv.f
[1,]  47 0.9555531 0.9745701
[2,]  48 0.9514781 0.9725746
[3,]  49 0.9472553 0.9696475
[4,]  50 0.9426571 0.9670942
[5,]  51 0.9373919 0.9638805
```

Expected lifetime II

```
> C50.m <- Surv.m[51:100]/Surv.m[51]
> C50.f <- Surv.f[51:100]/Surv.f[51]
> EL.m <- sum( C50.m * 1 )
> EL.f <- sum( C50.f * 1 )
> round( c(EL.m,EL.f), 2 )
[1] 27.90 31.44
```

Years of life lost

- — to diabetes, at age a
- the expected lifetime at age a if no DM
  minus
  the expected lifetime at age a if DM
- So you need to compute the (residual) expected lifetime (at age a):
  - for persons with DM
  - for persons without DM
- The same calculations, but now using age-specific mortality for diabetes patients...

Measures of this world

Short Course in Epidemiology
Advanced Stream
December 2016
IDEG, Vancouver, BC, Canada
http://BendixCarstensen.com/Epi/Courses/IDEG2015/

Cox

Inference in Multistate models

P.K. Andersen & N. Keiding
Interpretability and Importance of Functionals in Competing Risks and Multistate Models, *Stat Med, 2011* [1]:

1. Do not condition on the future
2. Do not regard individuals at risk after they have died
3. Stick to this world

Conditioning on the future

- . . . also known as "Immortal time bias", see *e.g.* S. Suissa: Immortal time bias in pharmaco-epidemiology, *Am. J. Epidemiol*, 2008 [2].
- Wrongly including persons' follow-up in the wrong state (namely the one reached some time in the future).
- Frequently caused by classification of **persons** instead of classification of **follow-up time**

Why these mistakes?

- Time is absent from survival analysis results
- Time is taken to be a **response** variable observed for each person
- Unit of analysis seems to be the person
- **Persons** classified by exposure
- The **real** unit of observation should be person-time

Time

- Time is a **covariate** — determinant of rates
- **Response** variable in survival / follow-up is bivariate:
  - **Differences** on the timescale (**risk** time, "exposure")
  - **Events**
- The relevant unit of observation is person-time:
  - small intervals of follow-up — "empirical rates"
  - (d_{it}, y_{it}): (event, (sojourn) time) for individual i at time t.
  - y is the **response** time, t is the **covariate** time
- Covariates relate to each interval of follow-up
- Allows **multiple** timescales, *e.g.* age and disease duration.

“Stick to this world”

In the paper by Andersen & Keiding this is primarily aimed at the use of “net survival”, that is the calculation of

$$\exp\left(-\int_0^t \lambda_c(s)\,\mathrm{d}s\right)$$

for a single cause of death
— formally for a non-exhaustive exit rate from a state.

Corresponds to the survival probability in the situation where:

1. all other causes of death are absent
2. the mortality, λ_c from cause c is unchanged

. . . which is indeed **not** of this world.

Sticking to this world

- Do not make predictions based on unrealistic assumptions:
  1. Mortality is 0
  2. Diabetes rates remain as now
- or
  1. Smallpox is eradicated
  2. . . . yet mortality remains the same
- I postulate a further specific feature of "this world":
- — it is **continuous**
- — in particular, death and disease rates vary **smoothly** by
  - age
  - calendar time
  - disease duration
  - . . .

A look at the Cox model

$$\lambda(t, x) = \lambda_0(t) \times \exp(x'\beta)$$

A model for the rate as a function of t and x.

The covariate t has a special status:

- Computationally, because all individuals contribute to (some of) the range of t.
- . . . the scale along which time is split (the risk sets)
- Conceptually it is less clear
  — t is but a covariate that varies within individual.
- Cox's approach profiles $\lambda_0(t)$ out.

The Cox-likelihood as profile likelihood

- One parameter per death time to describe the effect of time (i.e. the chosen timescale).

$$\log\big(\lambda(t, x_i)\big) = \log\big(\lambda_0(t)\big) + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} = \alpha_t + \eta_i$$

- Profile likelihood:
  - Derive estimates of α_t as function of data and βs — assuming constant rate between death times
  - Insert in likelihood, now only a function of data and βs
  - Turns out to be Cox's partial likelihood

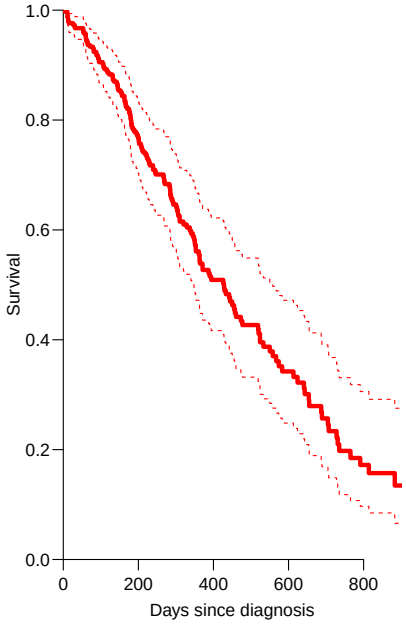
Splitting the dataset

- The Poisson approach needs a dataset of empirical rates (d, y) with suitably small values of y.
- — much larger than the original dataset
- — each individual contributes many empirical rates
- (one per risk-set contribution in Cox-modeling)
- From each empirical rate we get:
  - Poisson-response d
  - Risk time y
  - Covariate value for the timescale (time since entry, current age, current date, ...)
  - other covariates
- Modeling is by standard `glm` Poisson

Example: Mayo Clinic lung cancer

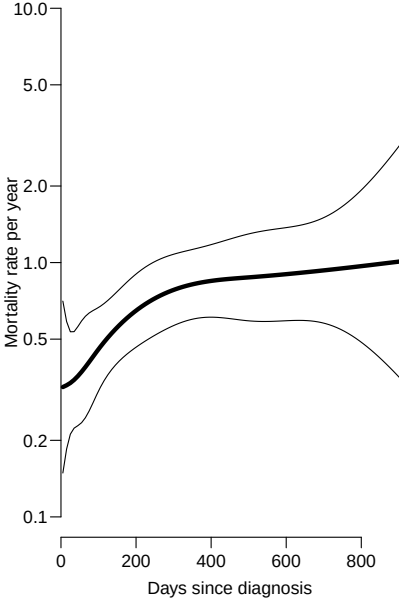
- Survival after lung cancer
- Covariates:
  - Age at diagnosis
  - Sex
  - Time since diagnosis
- Cox model
- Split data:
  - Poisson model, time as factor
  - Poisson model, time as spline

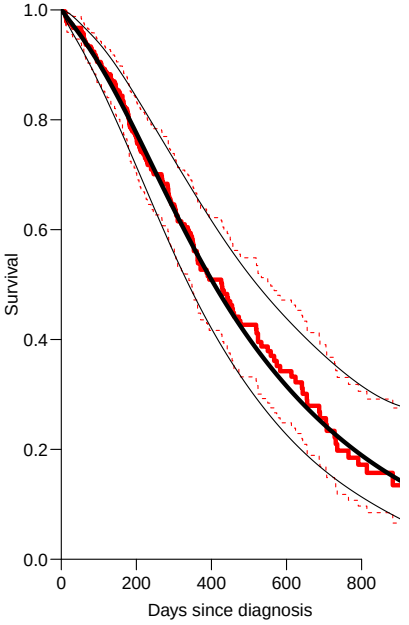
Mayo Clinic lung cancer 60 year old woman

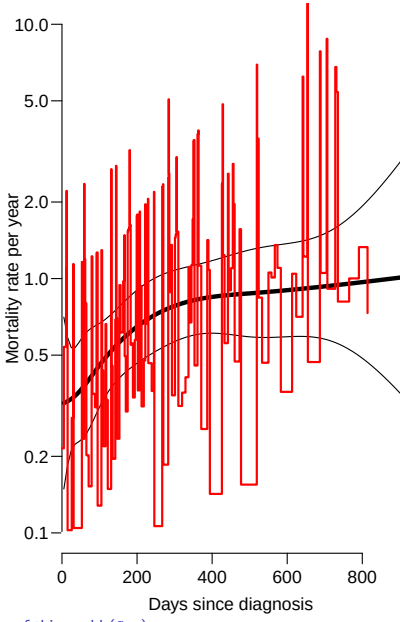


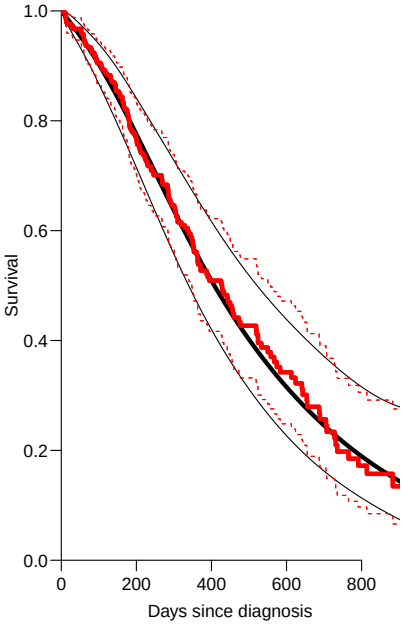
Example: Mayo Clinic lung cancer I

```
> round( cmp, 5 )
                    age    2.5%   97.5%     sex    2.5%   97.5%
Cox             1.01716 0.99894 1.03571 0.59896 0.43137 0.83165
Poisson-factor  1.01716 0.99894 1.03571 0.59896 0.43137 0.83165
Poisson-spline  1.01619 0.99803 1.03468 0.59983 0.43199 0.83287
```


10.0
5.0
2.0
1.0
0.5
0.2
0.1
Mortality rate per year
0
200
400
600
800
Days since diagnosis


1.0
0.8
0.6
0.4
0.2
0.0
Survival
0
200
400
600
800
Days since diagnosis


10.0
5.0
2.0
1.0
0.5
0.2
0.1
Mortality rate per year
0
200
400
600
800
Days since diagnosis


1.0
0.8
0.6
0.4
0.2
0.0
Survival
0
200
400
600
800
Days since diagnosis

```
> CM <- cbind( 1, Ns( seq(10,1000,10)-5, knots=t.kn ), 60, 1 )
> lambda <- ci.exp( mLs.pois.sp, ctr.mat=CM )
> Lambda <- ci.cum( mLs.pois.sp, ctr.mat=CM, intl=10 )[,-4]
> survP  <- exp(-rbind(0,Lambda))
```

What the Cox-model really is

Taking the life-table approach *ad absurdum* by:

- dividing time very finely and
- modeling one covariate, the time-scale, with one parameter per distinct value.
- $\Rightarrow$ difficult to access the baseline hazard
  — only the **cumulative hazard** (and survival) are accessible
- $\Rightarrow$ uninitiated tempted to show survival curves where irrelevant
- — or forgetting to show it where it is relevant

Modeling in this world

- Replace the α_ts by a parametric function $f(t)$ with a limited number of parameters, for example:
  - Piece-wise constant
  - Splines (linear, quadratic or cubic)
  - Fractional polynomials
- Brings model into "this world":
  - smoothly varying rates
  - parametric closed form representation of baseline hazard
  - finite no. of parameters
- Makes it really easy to extract and use the **hazard**:
  - to see how it looks
  - expected residual life time
  - state occupancy probabilities in multistate models
  - ...

Conclusions

Short Course in Epidemiology
Advanced Stream
December 2016
IDEG, Vancouver, BC, Canada
http://BendixCarstensen.com/Epi/Courses/IDEG2015/

conc

Conclusions

- Population based epidemiology requires:
  - Population level data (beware of sampling issues)
  - Population level models
- Models for hazard rates (incidence, mortality, morbidity)
- Ability to show and interpret the rates
- . . . using relevant derived measures:
  - RR relative to a given timepoint
  - Expected life time
  - etc.
- Survival models from clinical science almost always irrelevant.

References

P. K. Andersen and N. Keiding.
Interpretability and importance of functionals in competing risks and multistate models.
Stat Med, Nov 2011.

S. Suissa.
Immortal time bias in pharmaco-epidemiology.
Am. J. Epidemiol., 167:492–499, Feb 2008.