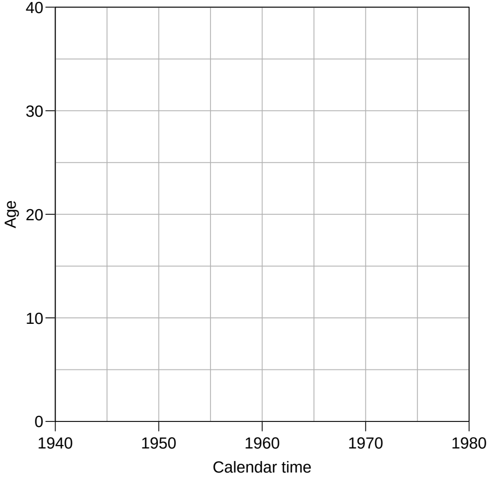
Statistical Analysis in the Lexis Diagram: Age-Period-Cohort models

Bendix Carstensen Steno Diabetes Center, Gentofte, Denmark
http://BendixCarstensen.com/

NSCE, Kellokoski, Finland
1 February 2014

www.bendixcarstensen.com/NSCE

Lexis diagram [1]

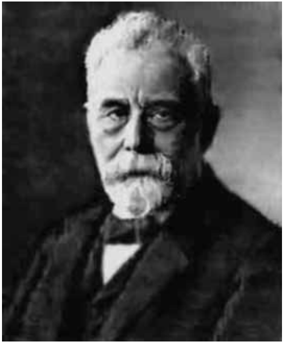


Disease registers record events.

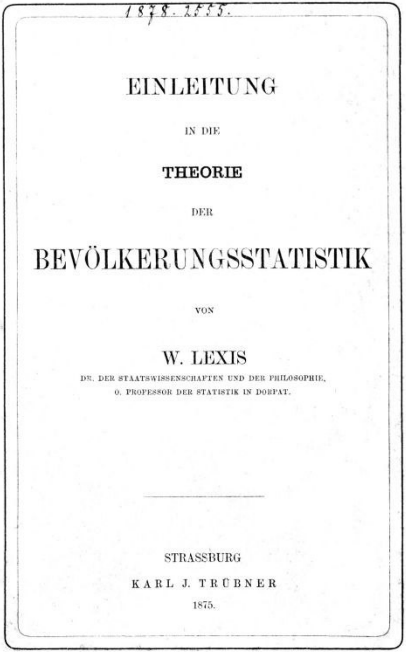
Official statistics collect population data.

[1] Named after the German statistician and economist **William Lexis** (1837–1914), who devised this diagram in the book "Einleitung in die Theorie der Bevölkerungsstatistik" (Karl J. Trübner, Strassburg, 1875).

Wilhelm Lexis



Wilhelm Lexis (1837–1914)
German statistician and economist.



1878. 2555.

EINLEITUNG

IN DIE

THEORIE

DER

BEVÖLKERUNGSSTATISTIK

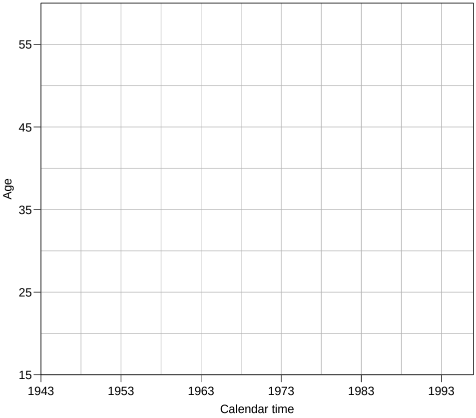
VON

W. LEXIS

DR. DER STAATSWISSENSCHAFTEN UND DER PHILOSOPHIE,
O. PROFESSOR DER STATISTIK IN DORPAT.

STRASSBURG
KARL J. TRÜBNER
1875.

Lexis diagram



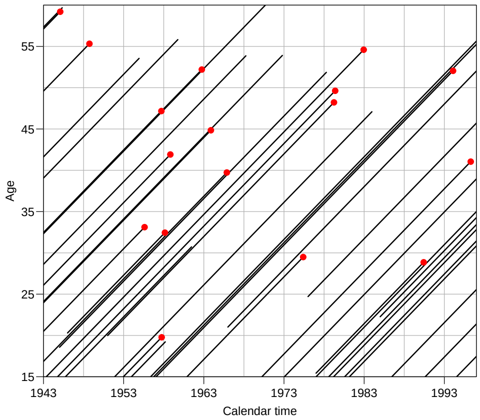
Registration of:

cases (D)

risk time,
person-years (Y)

in subsets of the
Lexis diagram.

Lexis diagram



Registration of:

cases (D)

risk time,
person-years (Y)

in subsets of the
Lexis diagram.

Rates available in
each subset.

Register data

Classification of **cases** (D_{ap}) by age at diagnosis and date of diagnosis, and **population** (Y_{ap}) by age at risk and date at risk, in compartments of the Lexis diagram, e.g.:

	Seminoma cases				Person-years			
Age	1943	1948	1953	1958	1943	1948	1953	1958
15	2	3	4	1	773812	744217	794123	972853
20	7	7	17	8	813022	744706	721810	770859
25	28	23	26	35	790501	781827	722968	698612
30	28	43	49	51	799293	774542	769298	711596
35	36	42	39	44	769356	782893	760213	760452
40	24	32	46	53	694073	754322	768471	749912

Reshape data to analysis form:

```
   A    P  D      Y
1 15 1943  2 773812
2 20 1943  7 813022
3 25 1943 28 790501
4 30 1943 28 799293
5 35 1943 36 769356
6 40 1943 24 694073
1 15 1948  3 744217
2 20 1948  7 744706
3 25 1948 23 781827
4 30 1948 43 774542
5 35 1948 42 782893
6 40 1948 32 754322
1 15 1953  4 794123
2 20 1953 17 721810
3 25 1953 26 722968
4 30 1953 49 769298
5 35 1953 39 760213
6 40 1953 46 768471
1 15 1958  1 972853
2 20 1958  8 770859
3 25 1958 35 698612
```

Tabulated data

Once data are in tabular form, models are restricted:

- Rates must be assumed constant in each cell of the table / subset of the Lexis diagram.
- With large cells it is customary to put a separate parameter on each level of the classifying factors.
- Output from the model will be rates and rate-ratios.
- Since we use multiplicative Poisson, usually the log rates and the log-RRs are reported

Register data - rates

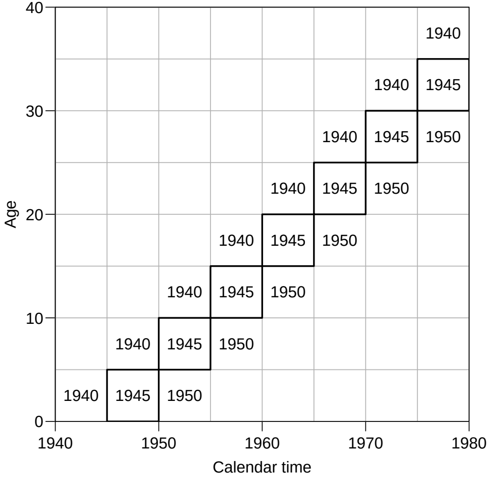
Rates in “tiles” of the Lexis diagram:

$$\lambda(a,p) = D_{ap}/Y_{ap}$$

Descriptive epidemiology based on disease registers:
How do the rates vary across by age and time?

- Age-specific rates for a given period.
- Age-standardized rates as a function of calendar time.
  (Weighted averages of the age-specific rates).

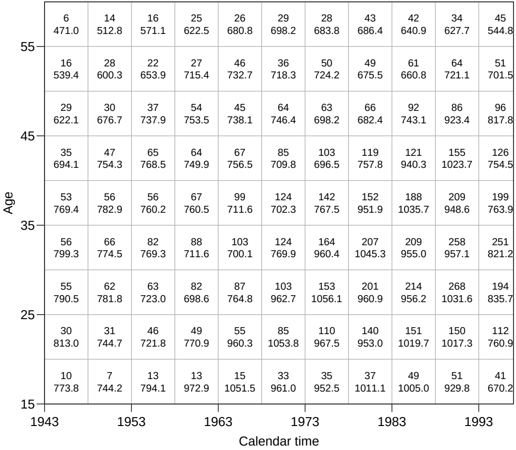
Synthetic cohorts



Events and risk time in cells along the diagonals are among persons with roughly same date of birth.

Successively overlapping 10-year periods.

Lexis diagram: data



Testis cancer cases in Denmark.

Male person-years in Denmark.

Data matrix: Testis cancer cases

Number of cases

	Date of diagnosis ($year - 1900$)								
Age	48–52	53–57	58–62	63–67	68–72	73–77	78–82	83–87	8
15–19	7	13	13	15	33	35	37	49	
20–24	31	46	49	55	85	110	140	151	
25–29	62	63	82	87	103	153	201	214	
30–34	66	82	88	103	124	164	207	209	
35–39	56	56	67	99	124	142	152	188	
40–44	47	65	64	67	85	103	119	121	
45–49	30	37	54	45	64	63	66	92	
50–54	28	22	27	46	36	50	49	61	
55–59	14	16	25	26	29	28	43	42	

Data matrix: Male risk time

1000 person-years

	Date of diagnosis ($year - 1900$)							
Age	48–52	53–57	58–62	63–67	68–72	73–77	78–82	83–87
15–19	744.2	794.1	972.9	1051.5	961.0	952.5	1011.1	1005.0
20–24	744.7	721.8	770.9	960.3	1053.8	967.5	953.0	1019.7
25–29	781.8	723.0	698.6	764.8	962.7	1056.1	960.9	956.2
30–34	774.5	769.3	711.6	700.1	769.9	960.4	1045.3	955.0
35–39	782.9	760.2	760.5	711.6	702.3	767.5	951.9	1035.7
40–44	754.3	768.5	749.9	756.5	709.8	696.5	757.8	940.3
45–49	676.7	737.9	753.5	738.1	746.4	698.2	682.4	743.1
50–54	600.3	653.9	715.4	732.7	718.3	724.2	675.5	660.8
55–59	512.8	571.1	622.5	680.8	698.2	683.8	686.4	640.9

Data matrix: Empirical rates

Rate per 1000,000 person-years

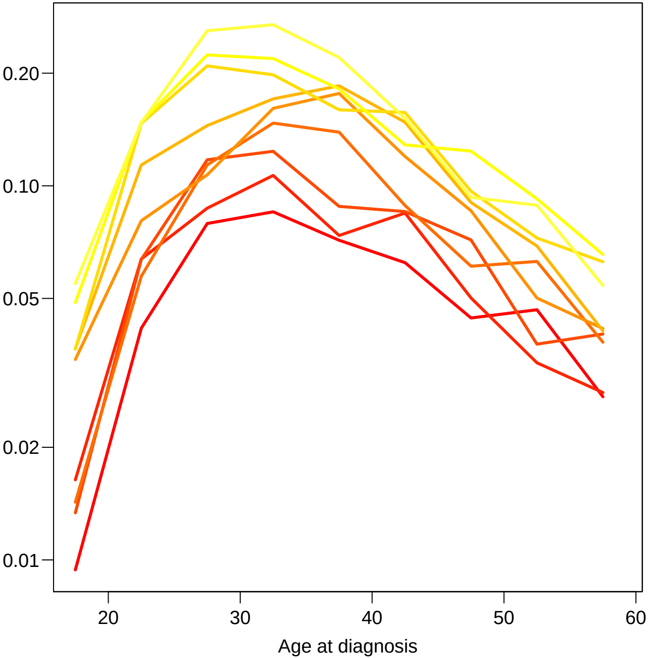
	Date of diagnosis $(year - 1900)$							
Age	48–52	53–57	58–62	63–67	68–72	73–77	78–82	83–87
15–19	9.4	16.4	13.4	14.3	34.3	36.7	36.6	48.8
20–24	41.6	63.7	63.6	57.3	80.7	113.7	146.9	148.1
25–29	79.3	87.1	117.4	113.8	107.0	144.9	209.2	223.8
30–34	85.2	106.6	123.7	147.1	161.1	170.8	198.0	218.8
35–39	71.5	73.7	88.1	139.1	176.6	185.0	159.7	181.5
40–44	62.3	84.6	85.3	88.6	119.8	147.9	157.0	128.7
45–49	44.3	50.1	71.7	61.0	85.7	90.2	96.7	123.8
50–54	46.6	33.6	37.7	62.8	50.1	69.0	72.5	92.3
55–59	27.3	28.0	40.2	38.2	41.5	40.9	62.6	65.5

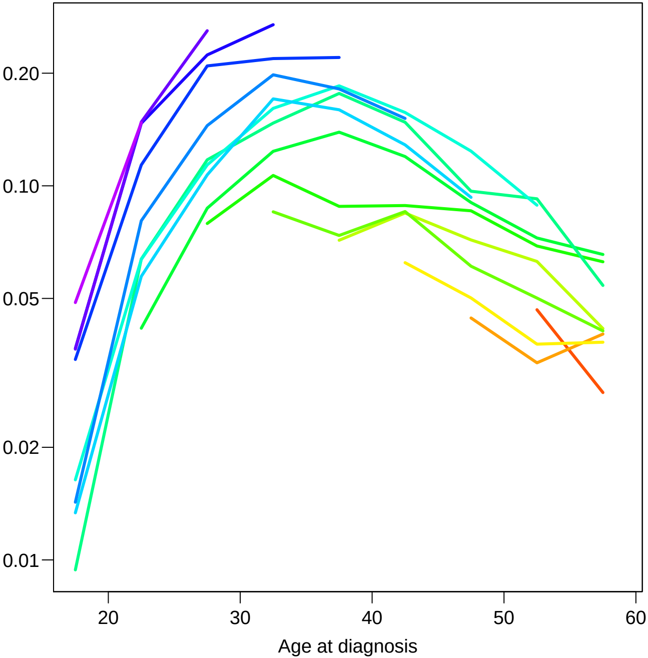
The classical plots

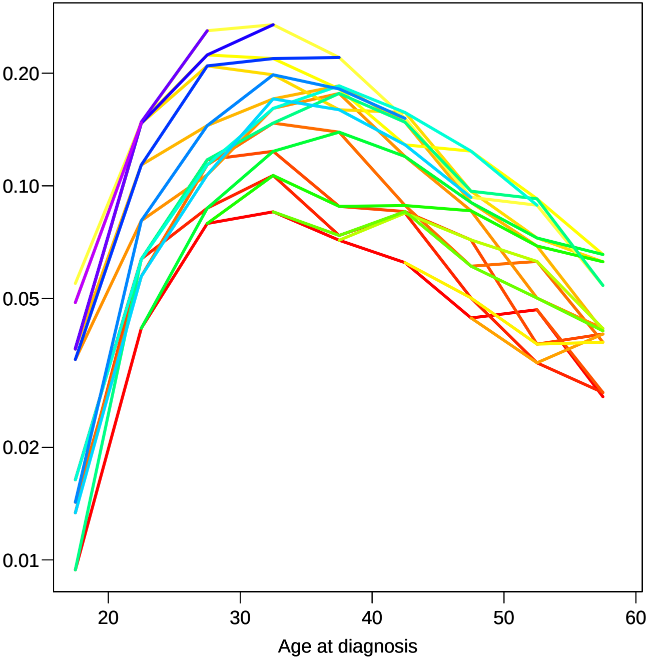
Given a table of rates classified by age and period, we can do 4 "classical" plots:

- Rates versus age at diagnosis (period):
  — rates in the same period connected.
- Rates versus age at diagnosis:
  — rates in the same birth-cohort connected.
- Rates versus date of diagnosis:
  — rates in the same ageclass connected.
- Rates versus date of date of birth:
  — rates in the same ageclass connected.

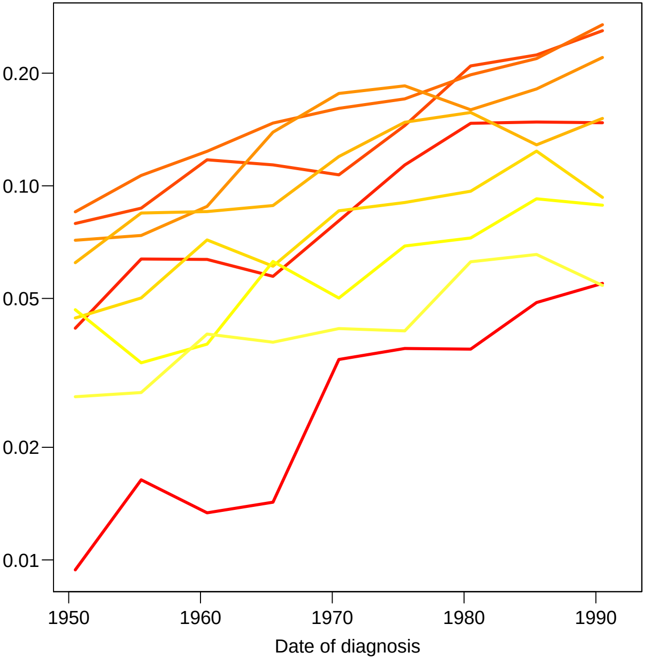
These plots can be produced by the R-function `rateplot`.


0.20
0.10
0.05
0.02
0.01
20
30
40
50
60
Age at diagnosis

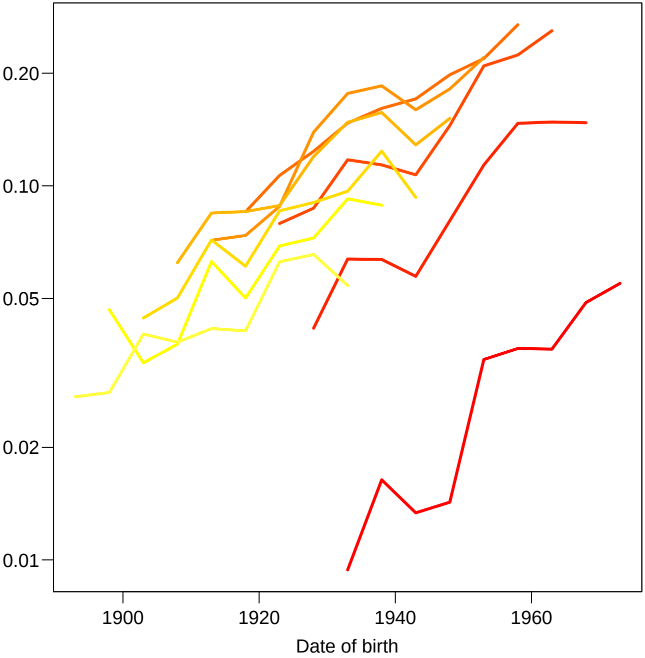

0.20
0.10
0.05
0.02
0.01
20
30
40
50
60
Age at diagnosis



0.20
0.10
0.05
0.02
0.01
20
30
40
50
60
Age at diagnosis



0.20
0.10
0.05
0.02
0.01
1950
1960
1970
1980
1990
Date of diagnosis


0.20
0.10
0.05
0.02
0.01
1900
1920
1940
1960
Date of birth

Age-period model

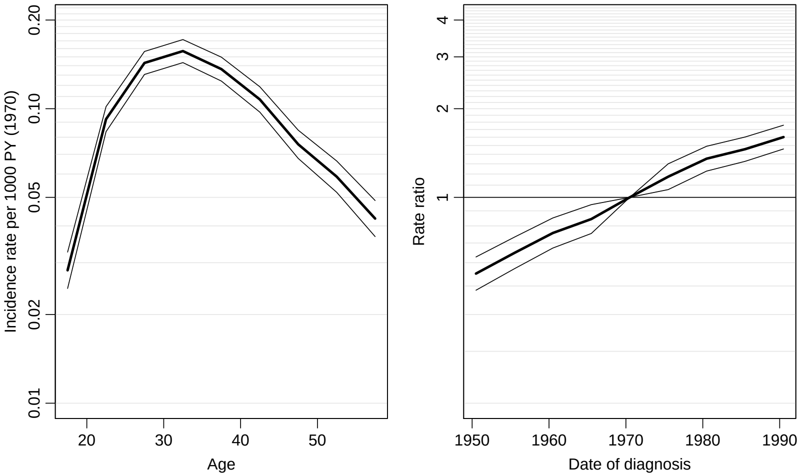
Rates are proportional between periods:

$$\lambda(a,p) = a_a \times b_p \qquad \text{or} \qquad \log[\lambda(a,p)] = \alpha_a + \beta_p$$

Choose p_0 as reference period, where $\beta_{p_0} = 0$

$$\log[\lambda(a,p_0)] = \alpha_a + \beta_{p_0} = \alpha_a$$

Estimates with confidence intervals



Age-cohort model

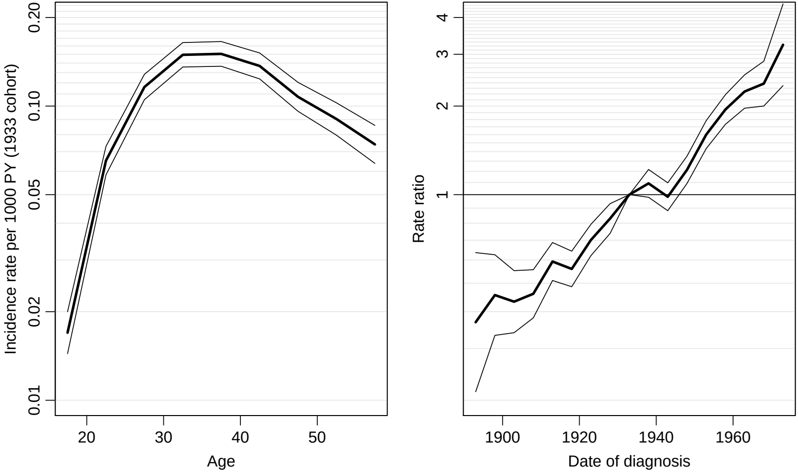
Rates are proportional between cohorts:

$$\lambda(a,c) = a_a \times c_c \qquad \text{or} \qquad \log[\lambda(a,p)] = \alpha_a + \gamma_c$$

Choose c_0 as reference cohort, where $\gamma_{c_0} = 0$

$$\log[\lambda(a,c_0)] = \alpha_a + \gamma_{c_0} = \alpha_a$$

Estimates with confidence intervals



Linear effect of period:

$$\log[\lambda(a,p)] = \alpha_a + \beta_p = \alpha_a + \beta(p - p_0)$$

that is, $\beta_p = \beta(p - p_0)$.

Linear effect of cohort:

$$\log[\lambda(a,p)] = \tilde{\alpha}_a + \gamma_c = \tilde{\alpha}_a + \gamma(c - c_0)$$

that is, $\gamma_c = \gamma(c - c_0)$

Age and linear effect of period:

```
> apd <- glm( D ~ factor( A ) - 1 + I(P-1970.5) +
+             offset( log( Y ) ),
+             family=poisson )
> summary( apd )

Call:
glm(formula = D ~ factor(A) - 1 + I(P - 1970.5) + offset(log(Y))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.97593  -0.77091   0.02809   0.95914   2.93076

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
factor(A)17.5 -3.58065    0.06306  -56.79   <2e-16
...
factor(A)57.5 -3.17579    0.06256  -50.77   <2e-16
I(P - 1970.5)  0.02653    0.00100   26.52   <2e-16

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 89358.53  on 81  degrees of freedom
Residual deviance:   126.07  on 71  degrees of freedom
```

Age and linear effect of cohort:

```
> acd <- glm( D ~ factor( A ) - 1 + I(C-1933) +
+             offset( log( Y ) ),
+             family=poisson )
> summary( acd )

Call:
glm(formula = D ~ factor(A) - 1 + I(C - 1933) + offset(log(Y)),

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.97593  -0.77091   0.02809   0.95914   2.93076

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
factor(A)17.5 -4.11117    0.06760  -60.82   <2e-16
...
factor(A)57.5 -2.64527    0.06423  -41.19   <2e-16
I(C - 1933)    0.02653    0.00100   26.52   <2e-16

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 89358.53  on 81  degrees of freedom
Residual deviance:   126.07  on 71  degrees of freedom
```

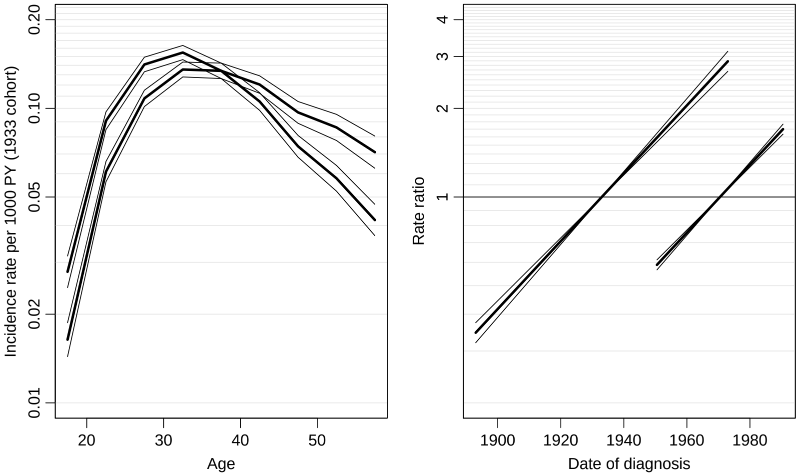
What goes on?

$$\begin{aligned}\alpha_a + \beta(p - p_0) &= \alpha_a + \beta\big(a + c - (a_0 + c_0)\big) \\ &= \underbrace{\alpha_a + \beta(a - a_0)}_{\text{cohort age-effect}} + \beta(c - c_0)\end{aligned}$$

The two models are the same.
The **parametrization** is different.

The age-curve refers either

- to a period (cross-sectional rates) or
- to a cohort (longitudinal rates).



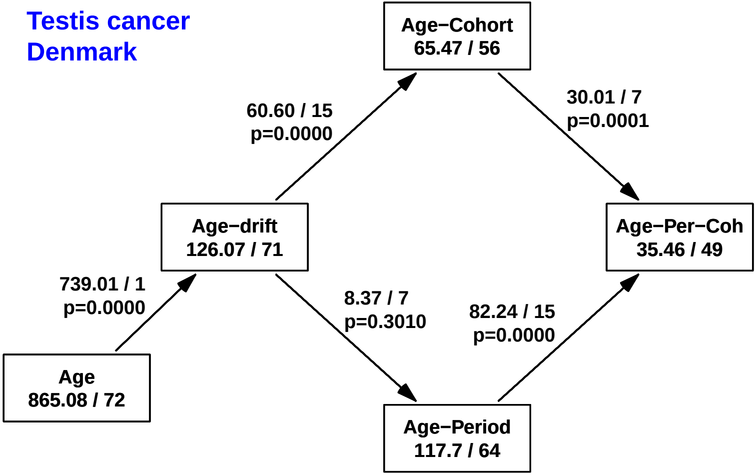
Which age-curve is period and which is cohort?

The age-period-cohort model

$$\log[\lambda(a,p)] = \alpha_a + \beta_p + \gamma_c$$

- Three effects:
  - Age (at diagnosis)
  - Period (of diagnosis)
  - Cohort (of birth)
- Modelled on the same *scale*.
- No assumptions about the *shape* of effects.

Relationship of models



Smooth functions

$$\log[\lambda(a, p)] = f(a) + g(p) + h(c)$$

Possible choices for parametric functions describing the effect of the three continuous variables:

- Polynomials / fractional polynomials.
- Linear / quadratic / cubic splines.
- Natural splines.

All of these contain the linear effect as special case,. . .

The identifiability problem still exists:

$$c = p - a \quad \Leftrightarrow \quad p - a - c = 0$$

$$\begin{aligned}
\lambda_{ap} &= f(a) + g(p) + h(c) \\
&= f(a) + g(p) + h(c) + \gamma(p - a - c) \\
&= f(a) - \mu_a - \gamma a + \\
&\quad\; g(p) + \mu_a + \mu_c + \gamma p + \\
&\quad\; h(c) - \mu_c - \gamma c
\end{aligned}$$

A decision on parametrization is needed.
It must be **external to the model**.

Parametrization of effects

There are still three “free” parameters:

$$
\begin{aligned}
\check{f}(a) &= f(a) - \mu_a \phantom{{}+\mu_c} - \gamma a \\
\check{g}(p) &= g(p) + \mu_a + \mu_c + \gamma p \\
\check{h}(c) &= h(c) \phantom{{}+\mu_a} - \mu_c - \gamma c
\end{aligned}
$$

Choose μ_a, μ_c and γ according to some criterion for the functions.

Parametrization principle

1. The age-function should be interpretable as log age-specific rates in cohort c_0 after adjustment for the period effect.
2. The cohort function is 0 at a reference cohort c_0, interpretable as log-RR relative to cohort c_0.
3. The period function is 0 on average with 0 slope, interpretable as log-RR relative to the age-cohort prediction. (residual log-RR).

Longitudinal or cohort age-effects.

Biologically interpretable — what happens during the lifespan of a cohort?

Implementation:

1. Obtain any set of parameters $f(a)$, $g(p)$, $h(c)$.
2. Extract the trend from the period effect:

$$\tilde{g}(p) = \hat{g}(p) - (\mu + \beta p)$$

3. Use the functions:

$$\begin{aligned}
\tilde{f}(a) &= \hat{f}(a) + \mu + \beta a + \hat{h}(c_0) + \beta c_0 \\
\tilde{g}(p) &= \hat{g}(p) - \mu - \beta p \\
\tilde{h}(c) &= \hat{h}(c) \qquad\quad + \beta c - \hat{h}(c_0) - \beta c_0
\end{aligned}$$

These functions fulfill the criteria.

How to?

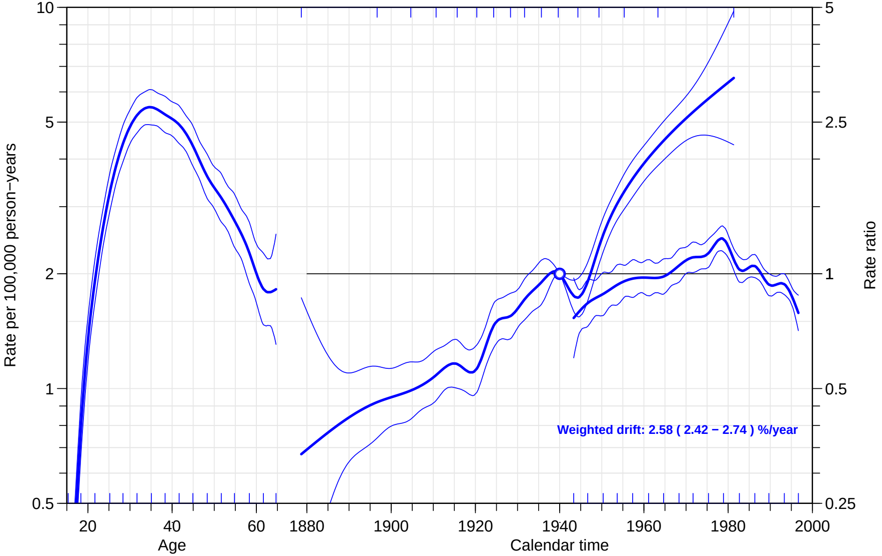
Implemented in `apc.fit`:

```
m1 <- apc.fit( A=lungDK$Ax,
               P=lungDK$Px,
               D=lungDK$D,
               Y=lungDK$Y/10^5,
           ref.c=1900 )
apc.plot( m1 )
```

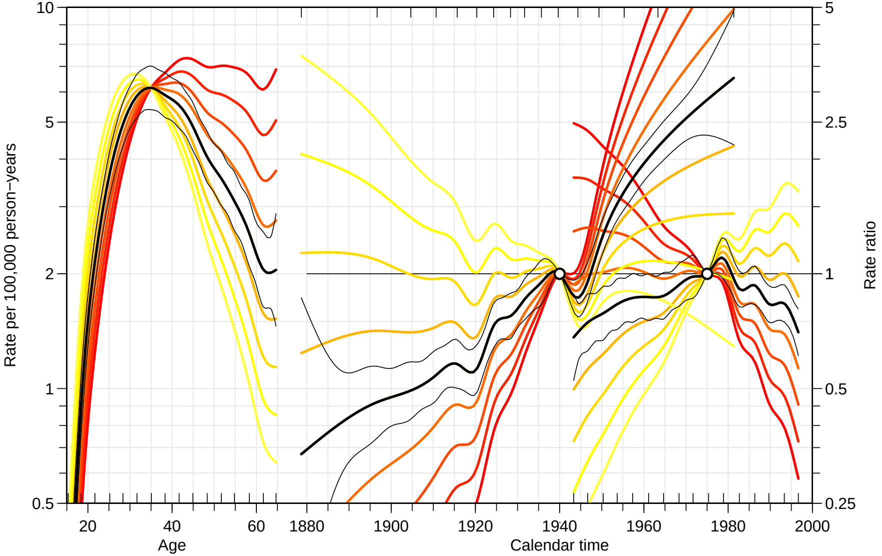
Consult the help page for details.


Rate
200
100
50
20
10
Rate ratio
2
1
0.5
0.2
0.1
Weighted i.p.
40
50
60
70
80
90
Age
1860
1880
1900
1920
1940
1960
1980
2000
Calendar time


200
100
50
20
10
Rate
2
1
0.5
0.2
0.1
Rate ratio
Naïve i.p.
Weighted i.p.
40
50
60
70
80
90
Age
1860
1880
1900
1920
1940
1960
1980
2000
Calendar time



10
5
2
1
0.5
Rate per 100,000 person−years
5
2.5
1
0.5
0.25
Rate ratio
20
40
60
Age
1880
1900
1920
1940
1960
1980
2000
Calendar time
Weighted drift: 2.58 (2.42 − 2.74) %/year



Rate per 100,000 person−years
10
5
2
1
0.5
20
40
60
Age
1880
1900
1920
1940
1960
1980
2000
Calendar time
5
2.5
1
0.5
0.25
Rate ratio

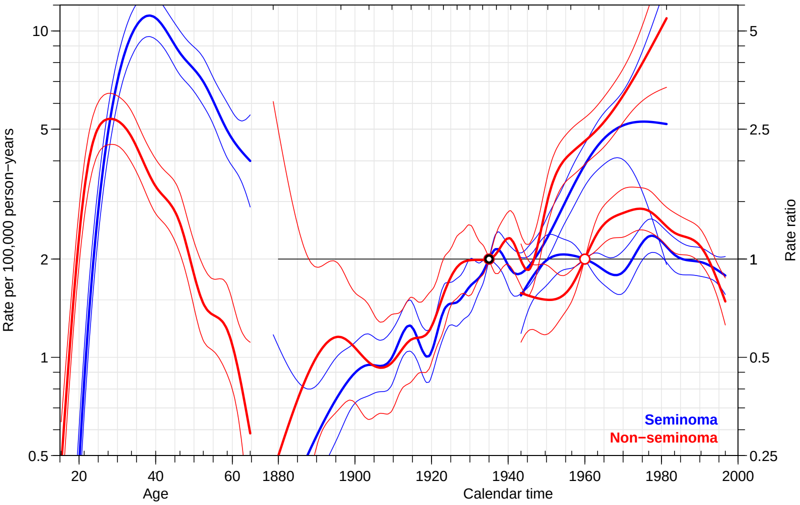
Two sets of data

Example: Testis cancer in Denmark, Seminoma and non-Seminoma cases.

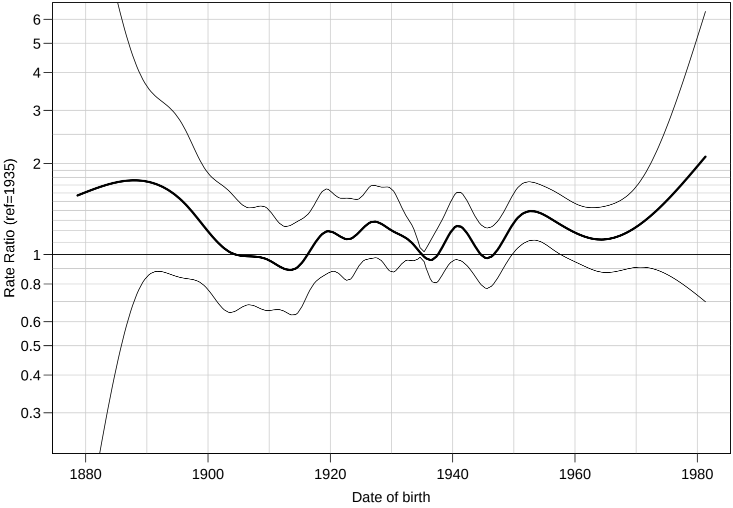
```
> stat.table( list( Histology=hist ),
+             list( D=sum(d), Y=sum(y/10^6) ),
+             margins = TRUE )
 ----------------------------
 Histology         D       Y
 ----------------------------
 1           4708.00  127.53
 2           3632.00  127.53
 3            466.00  127.53

 Total       8806.00  382.58
 ----------------------------
```

First step is separate analyses for each subtype.



Rate per 100,000 person−years
10
5
2
1
0.5
20
40
60
Age
1880
1900
1920
1940
1960
1980
2000
Calendar time
5
2.5
1
0.5
0.25
Rate ratio
Seminoma
Non−seminoma


6
5
4
3
2
1
0.8
0.6
0.5
0.4
0.3
Rate Ratio (ref=1935)
1880
1900
1920
1940
1960
1980
Date of birth

Rate Ratio (ref=1960)
6
5
4
3
2
1
0.8
0.6
0.5
0.4
0.3
1950
1960
1970
1980
1990
Date of diagnosis

Conclusions

- Categorization is a bad thing to do:
  - for data it's throwing away data
  - for modelling it's ignoring data
  - . . . or making silly assumptions
- **A**ge, **P**eriod and **C**ohort are **continuous** variables and should be treated as such:
- we want to see the continuous effct of these.
- Constraints needed **externally**,
- . . . just like it is needed to use a reference group if e.g. different occupational groups are compared.

Conclusions

- There is no solution to the identifiability problem,
- . . . only ways to cope with it.

Thanks for you attention.