6
Time

6.1 When do we start the clock?

In Chapter 5 we discussed the variation of rates with time. In that discussion, by assuming that all subjects entered the study at time zero, we implicitly interpreted time to mean time since entry into the study. However, there are many other ways of measuring time and some of these may be more relevant. For example, in epidemiology, it is usually important to consider the variation of rates with age, for which the origin is the date of birth, or with time since first exposure, for which the origin is the date of first exposure. Similarly, in clinical follow-up studies, time since diagnosis or start of treatment may be an important determinant of the failure rate. In different analyses, therefore, it may be relevant to start the clock at different points. Some possible choices for this starting point are described in Table 6.1.

6.2 Age-specific rates

Age is an extremely important variable in epidemiology, because the incidence and mortality rates of most diseases vary with age — often by several orders of magnitude. To ignore this variation runs the risk that comparisons between groups will be seriously distorted, or *confounded*, by differences in age structure.

The assumption that rates do not vary with age can be relaxed by dividing the age scale into bands and estimating a different *age-specific* rate in each band. If the follow-up period is short, so that the age of a

Table 6.1. Some time scales

Starting point	Time scale
Birth	Age
Any fixed date	Calendar time
First exposure	Time exposed
Entry into study	Time in study
Disease onset	Time since onset
Start of treatment	Time on treatment

Table 6.2. Entry and exit dates for the cohort of four subjects

Subject	Born	Entry	Exit	Age at entry	Outcome
1	1904	1943	1952	39	Lost
2	1924	1948	1955	24	Failure
3	1914	1945	1961	31	Study ends
4	1920	1948	1956	28	Unrelated death

subject does not change appreciably during follow-up, age-specific rates can be estimated by classifying subjects into age *groups* by their age at entry. Each subject appears in only one age group and a separate rate is estimated for each group. For longer studies it will be necessary to take account of changing age during the study, and to treat age properly — as a time scale. This scale is then divided into *bands* and a separate estimate of the rate is made within each age band as described in Chapter 5. In this latter analysis, a subject can pass through several age bands during the course of the study.

To see how the failures and observation time are divided between age bands consider the cohort of four subjects, shown in Table 6.2. Subject 1 is lost to follow-up in 1952, subject 2 fails in 1955, subject 3 is still under observation when the study period ends, and subject 4 dies from an unrelated cause in 1956. The date when a subject joins the cohort is called the entry date and the date when observation stops, for whatever reason, is called the exit date. The time between the entry and exit dates is the observation time for the subject. To simplify the exercises, we give dates only as years and will assume that all events take place on the first day of the year. In practice, times would be worked out as accurately as the data allow.

Exercise 6.1. What are the observation times for the members of this cohort?

Figure 6.1 shows the observation of the subjects in calendar time, while Figure 6.2 shows it on a scale where time is measured from each subject's date of birth. To estimate a rate for a particular age band the failures are allocated to the bands in which they occurred, and the observation time is divided according to how long the subjects spend in each of the age bands. For example, the age band 30-34, which is from exact age 30 to just less than exact age 35, contains one failure and 10 person-years of observation time, so the estimated rate is 1/10 per person-year.

In this example the observation times in the different time bands have been obtained from the figure, but in practice the total observation time in an age band is obtained by using the dates when the subject changes age bands. For example, subject 1 is 39 years old on entry so he starts in the age band 35–39. He changes age band in 1944 (when he is 40), and again in 1949 (when he is 45), and he leaves the study in 1952 (when he

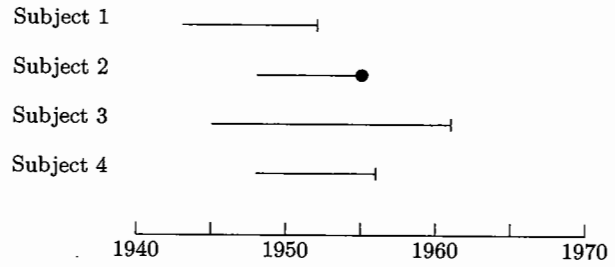


Fig. 6.1. Follow-up of four subjects by calendar time.

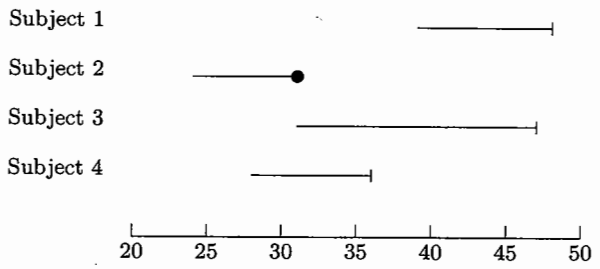


Fig. 6.2. Follow-up of four subjects by age.

emigrates). The observation time he spends in the different age bands is shown in Table 6.3.

As a check, the total observation time for subject 1 is from 1943 to 1952 which is 9 years, equal to the sum of the separate times spent in the different age bands.

Exercise 6.2. Subject 5 is born in 1931, joins the cohort in 1953, and is lost to follow-up in 1957. Divide the observation time for this subject between the five-year age bands shown in Figure 6.2.

Table 6.3. Time in each age band for subject 1

Age band	Date in	Date out	Time
35–39	1943	1944	1
40–44	1944	1949	5
45–49	1949	1952	3

Table 6.4. Woman-years and reference rates for a breast cancer study

Age	Woman-years	E & W rate per 100 000 woman-years
40–44	975	113
45–49	1079	162
50–54	2161	151
55–59	2793	183
60–64	3096	179

6.3 The expected number of failures

One reason for subdividing the total follow-up experience of a cohort into age bands is to determine whether the observed number of failures is more or less than we might have expected. Since mortality and incidence rates usually increase quite sharply with age, the distribution of person years observation between age bands is an extremely important determinant of the number of events we would expect to observe.

Table 6.4 shows the partition of woman-years between age bands for a cohort study of 974 women given a hormone treatment at menopause. During the follow-up period, 15 new cases of breast cancer occurred in the cohort. We might ask whether this is more or less than we would expect from national rates.

The third column of the table shows the age-specific incidence rates of breast cancer for England and Wales at the time the study was carried out. If the rates in the study population are the same as in the rest of England and Wales, the number of cases we would expect in each age band is simply the product of the woman-years observation and the rate. Thus, for the 40–44 age band, the expected number of cases is

$$975 \times \frac{113}{100\,000} = 1.10.$$

Exercise 6.3. Carry out these calculations for the remaining age groups and calculate the total expected number of cases of breast cancer.

This exercise shows that 16.77 cases are expected from national rates using the person years in the study. This expected number of cases is quite close to the observed 15, so that there is little suggestion that the rates in this cohort are unusual.

The expected number of cases, as calculated above, is not quite the same as the expected number in the usual statistical sense. The latter cannot depend upon the outcome of the study, but the former does, since the total person-time of observation in the study varies according to how many subjects fail and when. However, for the rare events studied by epidemiologists, this variation is small enough to be ignored.

6.4 Lexis diagrams

More than one time scale can be important in the same study. For example, mortality rates from cancer of the cervix depend upon age, as a result of the age-dependence of the incidence rate, and upon calendar time as a result of changes in treatment, population screening, and so on. The situation is further complicated by the strong dependence of the incidence of this disease upon sexual behaviour, which varies from one generation to the next.

The way to separate the effects of two time scales on a rate is to divide each scale into bands, usually of equal width, and to make a separate estimate of the rate for each pairing of bands. To see how this is done in practice it is best to show the subjects relative to the two scales simultaneously, in what is called a *Lexis diagram.*

The four subjects in Table 6.2 are shown relative to both age and calendar year simultaneously in the Lexis diagram in Figure 6.3. Each rectangular region in a Lexis diagram corresponds to a combination of two bands, one from each scale. To estimate rates for these combinations of bands the failures are allocated to the rectangles in which they occur and the observation time for each subject is divided between rectangles according to how long the subjects spends in each.

For example, subject 1 joins the cohort in 1943 aged 39. He changes age bands one year later in 1944 then 5 years later in 1949. He changes calendar periods in 1945 and 1950. Finally, observation stops in 1952. The subdivision of the observation time for this subject between different age and calendar period combinations is shown in Figure 6.4. Note that the times in the different bands add to 9 years, the total observation time for this subject. For each combination of age band and calendar period the rate is estimated by dividing the number of failures by the person-time of observation.

Exercise 6.4. Trace the progress of subject 1 through the squares in Figure 6.3 and verify the results given above. Divide the observation time for subject 2 between combinations of five-year bands of age and calendar time in the same way.

The same procedure can be used to separate the effect of age from the effect of time since entry, although there may not be enough data for some combinations of age and time since entry to estimate a rate. Figure 6.5 shows the four subjects in the cohort relative to age and time since entry. Five-year bands have again been chosen for both scales.

Exercise 6.5. Divide the observation time for subject 1 between different combinations of five-year bands of age and time since entry.

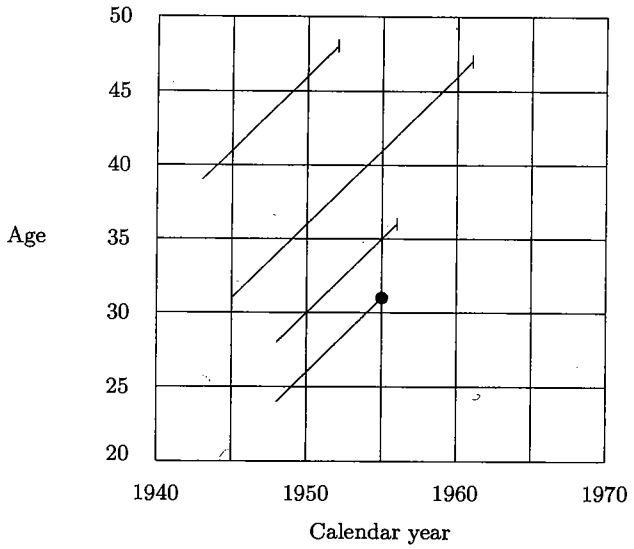


Fig. 6.3. Lexis diagram showing age and calendar period.

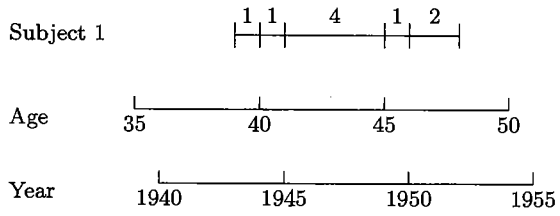


Fig. 6.4. Follow-up of subject 1 by age and calendar time.

6.5 Reference rates by calendar period

Reference rates, used to calculate the expected numbers of failures, usually come from national rates tabulated by age, sex, and calendar period. In the UK these are calculated using an approximate figure for the person-years. For example, the all-cause mortality rate for the age band 50–54 during 1983 is estimated by D/Y where D is the number of deaths during 1983 for which the subject's age at death was in the range 50–54, and Y is the person-time lived during 1983 by that part of the population whose ages were in the range 50–54 during 1983. Since the exact value of Y is not

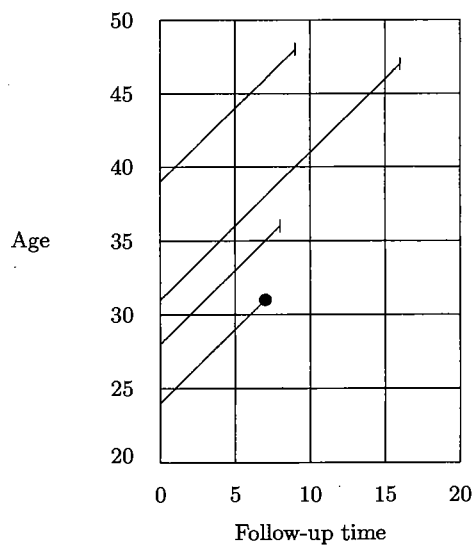


Fig. 6.5. Lexis diagram showing age and time since entry.

known an approximate value is obtained from

$$Y \approx \text{Population aged 50–54 in mid-1983} \times 1 \text{ year}.$$

For five-year calendar periods such as 1981–85,

$$Y \approx \text{Population aged 50–54 in mid-1983} \times 5 \text{ years}.$$

The population in the different age bands for any year is obtained from the census; directly for census years and indirectly for inter-census years by updating the last census by births, deaths, and migration.

Exercise 6.6. The total number of deaths from cancer of the lung in the SW region of England during the years 1981–88 were males: 14 751, females: 5420. The 1984 population of the region is estimated to be males: 2 154 900, females: 2 306 300. Calculate the mortality rate per 10^6 person-years for males and females separately.

When follow-up of a cohort takes place over an extended calendar period, the national age-specific rates will usually vary over this period, making it difficult to choose a single set of age-specific rates to use for comparison purposes. The solution is to compute the expected number of events by both age and calendar period, using the appropriate national rates for each calendar time period. To do this the person-years observation in the co-

Table 6.5. Mortality following X-irradiation

Cause of death	Number of deaths		Ratio
	Observed, D	Expected, E	D/E
Cancers:			
Leukaemia	31	6.47	4.79
Colon	28	17.30	1.62
Heavily irradiated sites	259	167.50	1.55
Lightly irradiated sites	79	65.65	1.20
All neoplasms	397	256.92	1.55
Other causes	1362	804.68	1.69
All causes	1759	1061.61	1.66

hort study must be partitioned by age and calendar period. The expected number of failures can then be calculated for each combination of age and calendar period, as before, by multiplying the person-years observation by the appropriate national rate. Addition over all combinations of age and calendar period yields an expected number of cases which takes account of variation in national rates with both age and calendar time.

An example of this kind of calculation appears in Table 6.5, which shows some results taken from a study of cancer mortality in a cohort of ankylosing spondilitis patients who had been treated with a single course of X-irradiation of the spine.* The follow-up of each patient started in the year of treatment (1935–1954) and continued until death, migration or 1970 (the date when this analysis was carried out). Follow-up was also terminated by a second course of treatment because the aim was to study the effect of a single course of X-rays and the time before this effect became apparent. The study was carried out in Great Britain and Northern Ireland, and the expected numbers of deaths calculated using the national rates for England and Wales, tabulated by five-year bands for both age and calendar time. It can be seen that mortality from all causes was higher in this cohort than in the reference population. Although accounting for relatively few excess deaths, the *ratio* of observed to expected deaths was particularly high for leukaemia. This ratio is an important index in epidemiology and is called the *standardized mortality ratio* (SMR). We shall discuss it further in Chapter 15.

Exercise 6.7. Table 6.6 subdivides the observed and expected deaths from leukaemia according to time since X-ray treatment. How would this table have been calculated?

*From Smith, P.G. and Doll, R.(1982) *British Medical Journal*, **284**, 449–460.

Table 6.6. Leukaemia deaths by time since treatment

	Time since treatment (years)							
	0–2	3–5	6–8	9–11	12–14	15–17	18–20	>20
Observed	6	10	6	3	1	4	1	0
Expected	1.00	0.89	0.87	0.90	0.96	0.90	0.55	0.40
Ratio	6.00	11.24	6.90	3.33	1.04	4.44	1.82	0.00

Solutions to the exercises

6.1 The observation times for the four subjects are 9, 7, 16, and 8 years respectively.

6.2 Subject 5 is 22 years of age on joining the cohort and 26 when lost to follow-up. She contributes 3 years to the band 20–24, and 1 year to the band 25–29.

6.3 The expected numbers of cases in the five age bands are 1.10, 1.75, 3.26, 5.11, and 5.54. The sum of these values is 16.76, but working to full accuracy we obtain 16.77 for the total expected number of cases.

6.4 The Age×Period bands in which subject 2 was observed are as follows:

Age	Calendar period	Time in band
20–24	1945–49	1
25–29	1945–49	1
25–29	1950–54	4
30–34	1950–54	1

6.5 The Age×Follow-up bands in which subject 1 was observed are as follows:

Age	Follow-up time	Time in band
35–39	0–4	1
40–44	0–4	4
40–44	5–9	1
45–49	5–9	3

6.6 The estimated rate for males is

$$\frac{14\,751}{2\,154\,900 \times 8} = 856 \text{ per } 10^6 \text{ person-years}$$

and the estimated rate for females is

$$\frac{5\,420}{2\,306\,300 \times 8} = 294 \text{ per } 10^6 \text{ person-years.}$$

6.7 The follow-up of each subject can be represented by a line on a three-dimensional Lexis diagram with axes: age, period, and time since treatment. Age and period were divided into five-year bands and time since treatment into three-year bands. Observed deaths and person-years can be assigned to cells in the resulting three-dimensional table. Multiplication of person-years by national rates gives the expected number of deaths for each cell. Table 6.6 is formed by adding this table over age and period.

6 Time

6.1 When do we start the clock?

Examples JH has dealt with include the analysis of longevity of

- The Titanic survivors, where the two time scales are (i) age (years elapsed since birth) and (ii) 'survivor-time', the years elapsed since the April 15, 1912 sinking;
- Oscar nominees, where the two time scales are (i) age and (ii) nominee-time', the years elapsed since first being nominated for an Oscar;
- Nobel Prize nominees, where the two time scales are (i) age and (ii) 'nominee-time', the years elapsed since first being nominated for a Nobel Prize;
- Jazz musicians, where the two time scales are (i) age and (ii) performer-time', the years elapsed since first becoming a jazz musician;
- Popes versus artists;
- Baseball Hall of Famers versus players who were nominated by not inducted;
- Rock Stars who become famous early versus later (or not at all).

  For more details on these examples, see `bios601/CandHchapter06/`

For more on the choice of time scale, Google "Multiple time scales in survival analysis." or find the articles that cite the 1979 Applied Statistics article by Farewell and Cox "*A note on multiple time scales in life testing.*"

There is also the interesting article *The two-way proportional hazards model* by Efron in J. R. Statist. Soc. B (2002) 64, Part 4, pp. 899-909, applied to "patient histories in a study of heart transplant recipients treated at the Stanford Medical Center between 1980 and 1996; some 110 of the patients suffered a *serious bacterial infection*, their infection times ranging from a few *days after transplantation* to nearly 9 years, these being the observed lifetimes that would usually be featured in a proportional hazards analysis of the infection process. In this case, however, the investigators' *main interest centred on calendar date*: was the *incidence rate* of bacterial infections *declining over the course of the study?* Incidence is itself a hazard rate, in the simplest situation the number of new cases per eligible subject per unit time, and it is natural to answer the question with a hazard rate analysis."

6.2 Age-specific rates

"*To ignore this variation [of incidence and mortality rates with age] runs the risk that comparisons between groups will be seriously distorted, or confounded, by differences in age structure.*"

It's good to have a few handy real examples of *age-confounding* that are easily understood by non-statisticians. Two immediately come to mind (i) the overall death rate is higher in Canada than Ethiopia (ii) the higher death rate among non-smokers in a 20-year follow-up study of smokers and non-smokers [Does Smoking Improve Survival? `www.whfreeman.com/statistics/ips/eesee4/eesees4.htm`; this is also described in chapter 1 of Rothman 2002, with finer age-categories]

"*For longer studies it will be necessary to take account of changing age during the study, and to treat age properly - as a time scale. This scale is then divided into bands and a separate estimate of the rate is made within each age band as described in Chapter 5. In this latter analysis, a subject can pass through several age bands during the course of the study.*"

Not only can a subject pass through several age bands but she can also change from one 'exposure' category to another – as in the Oscars exercise.

6.3 The expected number of failures

"*One reason for subdividing the total follow-up experience of a cohort into age bands is to determine whether the observed number of failures is more or less than we might have expected. Since mortality and incidence rates usually increase quite sharply with age, the distribution of person years observation between age bands is an extremely important determinant of the number of events we would expect to observe.*"

It is not clear what is the basis for the "expectation" i.e., whether it is a 'what if' comparison against *external* rates, or an *internal* one against the rates in a comparison group constructed and followed by the investigators. One can think of the 'expected number' of 16.77 cases in exercise 6.3 as the number one would expect in a scaled-down version of England and Wales (E&W), scaled down to the same sample size (974 women) followed for the same cell-specific numbers of person years as those shown in Table 6.4. In other words, it as as thought one had

974 treated by HRT	974 from E&W, same age & follow-up, untreated
15 cases	16.7 cases

Of course, the fact that the 16.7 is based on observed rates in the whole of

E&W means that it is not subject to the same degree of random variation as is the number of cases in the actual cohort. With this solid a basis for it, the expected number is usually taken to be a constant, so only one standard error (SE) is involved in the 15 vs. 16.7 comparison – the one associated with the 15.

"The expected number of cases, as calculated above, is not quite the same as the expected number in the usual statistical sense. The latter cannot depend upon the outcome of the study, but the former does."

C&H are saying that the numbers of Woman-years in the second column of Table 6.4 are *random variables*: they would not have been known ahead of time. For some 15 women – the 15 being a random variable – the follow-up was terminated by the event of interest. Likewise, any terminations for other reasons might also be unpredictable ahead of time. However, if these are not related to the person's probability of a future event, they don't have a great influence on the sampling behaviour of the estimators of interest.

6.4 Lexis diagrams

`en.wikipedia.org/wiki/` Wilhelm Lexis (1837-1914) was an eminent German statistician, economist, and social scientist and a founder of the interdisciplinary study of insurance.

The "Lexis diagram", in which lifelines are displayed as 45-degree lines on a grid with age on the vertical axis and calendar year on the horizontal axis, is very helpful in epidemiology, and in survival analysis with 2 time scales.

The `Epi` package for `R` has several functions that make it easy to convert the data of the type shown in Table 6.2 into the person-year segments shown Figure 6.3. Previously, this was a very laborious computing process.

Once we have the tabulated person years and cases in each Lexis rectangle (the cells don't have to be square), we can calculate the expected number of cases if a specified set of external rates applied, or make internal rectangle-by-rectangle comparisons, and thus a summary of these comparisons. We can also use them to fit (Poisson) regression models for rates.

Here is the `R` code, and some of its output, for the data in C&H Table 6.2.

```
library(Epi)

id = c(1,2,3,4);
yr.birth = c(1904,1924,1914,1920);
yr.entry = c(1943,1948,1945,1948);
yr.exit  = c(1952,1955,1961,1956);
fail = c(0, 1, 0, 0) );

ds=data.frame(id, yr.birth, yr.entry, yr.exit, fail); ds

  id yr.birth yr.entry yr.exit fail
1  1     1904     1943    1952    0
2  2     1924     1948    1955    1
3  3     1914     1945    1961    0
4  4     1920     1948    1956    0


# Define as Lexis object with timescales calendar time and age

Lexis <- Lexis( entry = list( calendar.year = yr.entry ),
                exit  = list( calendar.year = yr.exit, age = yr.exit - yr.birth ),
         exit.status  = fail,
                data  = ds )
Lexis

  calendar.year age lex.dur lex.Cst lex.Xst lex.id id yr.birth yr.entry yr.exit fail

1          1943  39       9       0       0      1  1     1904     1943    1952    0
2          1948  24       7       0       1      2  2     1924     1948    1955    1
3          1945  31      16       0       0      3  3     1914     1945    1961    0
4          1948  28       8       0       0      4  4     1920     1948    1956    0

# Default plot of follow-up

plot(Lexis)

# With a grid and deaths as endpoints

plot(Lexis, grid=0:5*5, col="black" )
points(Lexis, pch=c(NA,16)[Lexis$lex.Xst+1] )

# With a lot of bells and whistles:  [ *** SEE PLOT NEXT PAGE *** ]

plot(Lexis, grid=0:20*5, col="black", xaxs="i", yaxs="i",
      xlim=c(1940,1965), ylim=c(20,50), lwd=3, las=1 )
points(Lexis, pch=c(NA,16)[Lexis$lex.Xst+1], col="red", cex=1.5 )

# Split time along two time-axes

L2 = splitLexis(Lexis,breaks=seq(1940,1965,5),
   time.scale="calendar.year")
L2 = splitLexis(L2,   breaks=seq(20,50,5), time.scale="age" )
str( L2 )
```

```
L2
```

	lex.id	calendar.year	age	lex.dur	lex.Cst	lex.Xst	id	yr.birth	yr.entry	yr.exit	fail
1	1	1943	39	1	0	0	1	1904	1943	1952	0
2	1	1944	40	1	0	0	1	1904	1943	1952	0
3	1	1945	41	4	0	0	1	1904	1943	1952	0
4	1	1949	45	1	0	0	1	1904	1943	1952	0
5	1	1950	46	2	0	0	1	1904	1943	1952	0
6	2	1948	24	1	0	0	2	1924	1948	1955	1
7	2	1949	25	1	0	0	2	1924	1948	1955	1
8	2	1950	26	4	0	0	2	1924	1948	1955	1
9	2	1954	30	1	0	1	2	1924	1948	1955	1
10	3	1945	31	4	0	0	3	1914	1945	1961	0
11	3	1949	35	1	0	0	3	1914	1945	1961	0
12	3	1950	36	4	0	0	3	1914	1945	1961	0
13	3	1954	40	1	0	0	3	1914	1945	1961	0
14	3	1955	41	4	0	0	3	1914	1945	1961	0
15	3	1959	45	1	0	0	3	1914	1945	1961	0
16	3	1960	46	1	0	0	3	1914	1945	1961	0
17	4	1948	28	2	0	0	4	1920	1948	1956	0
18	4	1950	30	5	0	0	4	1920	1948	1956	0
19	4	1955	35	1	0	0	4	1920	1948	1956	0

```
# Tabulate the cases and the person-years

summary( L2 )

tapply( status(L2,"exit")==1, list( timeBand(L2,"age","left"),
          timeBand(L2,"calendar.year","left") ), sum )
```

	1940	1945	1950	1955	1960
20	NA	0	NA	NA	NA
25	NA	0	0	NA	NA
30	NA	0	1	NA	NA
35	0	0	0	0	NA
40	0	0	0	0	NA
45	NA	0	0	0	0

```
tapply( dur(L2),  list( timeBand(L2,"age","left"),
          timeBand(L2,"calendar.year","left") ), sum )
```

	1940	1945	1950	1955	1960
20	NA	1	NA	NA	NA
25	NA	3	4	NA	NA
30	NA	4	6	NA	NA
35	1	1	4	1	NA
40	1	4	1	4	NA
45	NA	1	2	1	1

```
> summary( L2 )

Transitions:
     To
From  0 1 Records:  Events:  Risk time:
   0 18 1        19         1           40

Rates:
     To
From 0    1 Total
   0 0 0.02  0.02
```

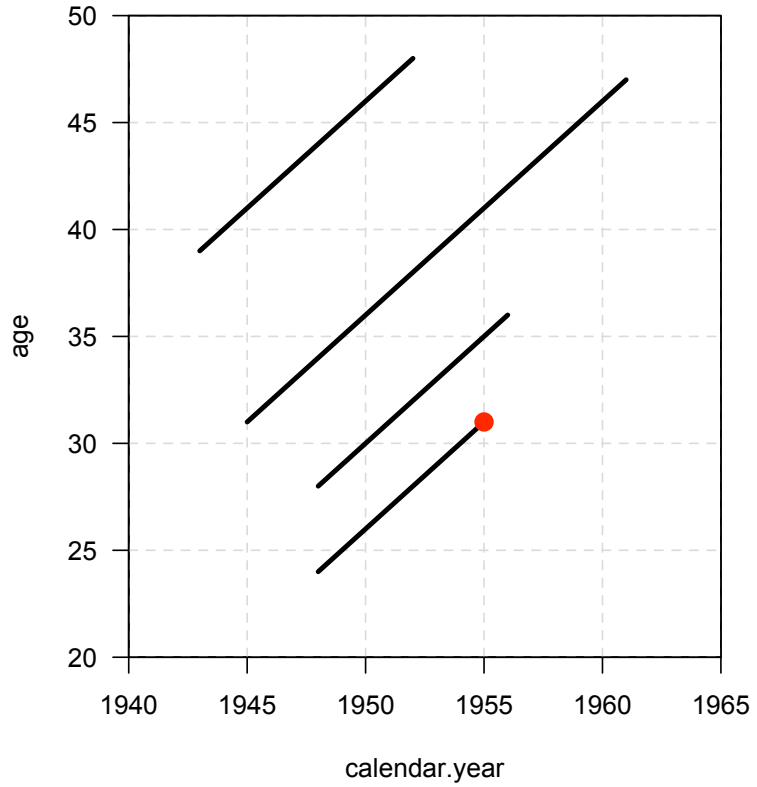


Figure 1: Lexis Diagram, from `Epi` package in `R`

Supplementary Exercise 6.1. Death rates in those who survived the sinking of the Titanic vs. in the sex-and age-matched US general population, together with some other investigations

Under the heading Longevity Comparisons in the resources for C&H chapter 06 you will find links to (a) the Titanic longevity data set (b) USA death rates (within 5 x 5 rectangles, called '*quinquinquennia*') from the Berkeley Mortality Database.[1] You will also find some R code that uses the Epi package to create – for each passenger – the durations in and exit status from each quinquinquennium, then aggregates these over all the persons traversing each quinquinquennium, etc.

1. Convert each survivor's record into the experience in the (age, period) quinquinquennia traversed, i.e the number of years spent in the rectangle, and the status (e.g., $d = 0$ if alive, 1 if dead) at the end of these years. Rather than program the calculations from scratch, two possibilities are `http://epi.klinikum.uni-muenster.de/pamcomp/pamcomp.html` – which some people used last year – and the `R` 'Epi' package `http://staff.pubhealth.ku.dk/~bxc/Epi/` The key functions in the latter are `Lexis` (and associated plotting functions) and `splitLexis`, which, when applied twice, calculates the time spent, and exit status from each quinquinquennium. The 'bogus example' in the documentation of the `splitLexis` function illustrates these, while the example on the notes for C&H chapter 6 shows the application to the 4-person cohort used in that chapter.

2. How much higher/lower is the *set* of age-specific death rates for male Titanic survivors than that for the general US population? for female survivors? Answer in two ways: first, calculate sex-specific observed/expected ratios, where the numerator is the total number of deaths observed in the sex-specific cohort, and the denominator is the sum of the expected numbers of deaths in these cells, using the USA age-sex-period death rates; second, calculate sex-specific Mantel-Haenszel summary incidence ratios (Rothman terminology) or incidence density ratios (Miettinen terminology) or mortality rate ratios (everyone's terminology), using age and period as 'strata.'[2] Assume that each of the USA death rates is based on a denominator of one million person years.[3] Assume that the death rates after 1995 are the same as those in 1990-95.

3. 'On average,' [4], for the age-span 40-90 in the period 1990-1995, how much higher are the USA age-specific male death rates in males than females? Answer by plotting the log of the male:female death rate ratio vs age, (or the two separate sets of log-death-rates on the same graph), and taking some 'typical' value for the ratio. Are you comfortable giving a single ratio? i.e., is the mortality-rate-ratio (M:F) reasonably constant over that age-span?

4. The previous question refers to cross-sectional rates, i.e., those in a specified *period*.[5] On average, over the age-span 40-90 in the 1900 *birth-cohort*, how much higher are the USA age-specific death rates in males than females? Answer by plotting the log of the male:female death rate ratio vs age, (or the two separate sets of log-death-rates on the same graph), and taking some 'typical' value for the ratio. Are you comfortable giving a single ratio? i.e., is the mortality-rate-ratio (M:F) *reasonably* constant over that age-span?

5. For the age-span 40-90, in a single number describe how much age-and specific death rates have fallen over the 20th century (the changes may be more subtle that this, so your answer will necessarily be a simplification).

6. For the Titanic survivors, was there a gradient in mortality rates across the 3 passenger classes?

Supplementary Exercise 6.2. Mortality of performers while in the 'still hoping to win' vs in the 'already a winner' state[6]

1. Divide the performer-years into those spent as Oscar nominees and as Oscar winners and then subdivide these into quinquinquennia.

2. Compare the death rates in the performer-years spent as nominees versus those spent as winners. Do so using both 'adjusted' expected numbers and purely-internal comparisons.

[1.] This site, `http://www.demog.berkeley.edu/~bmd/index.html`, contains historical lifetable and death rate data for the USA and other countries.

[2] As is illustrated in equation 8-5 in Rothman 2002, the formula is

$$\frac{\sum_{strata}(no.\ of\ cases,\ index\ category) \times (py,\ ref.\ category)/(py\ in\ stratum)}{\sum_{strata}(no.\ of\ cases,\ ref.\ category) \times (py,\ index\ category)/(py\ in\ stratum)}$$

[3] If the ratio of the amount of experience in the ref. category to that in the index category goes to infinity, the M-H summary ratio converges to $\sum_{strata} O / \sum_{strata} E = O/E$.

[4] Even if the average is not representative.

[5] Cross-sectional rates are what are used to make 'current' or 'period' lifetables, by far the more common type of lifetable.

[6] The link to the Oscars material can be found under the heading Longevity Comparisons in the resources for C&H chapter 06. Once on the c634 page, scroll down to the longevity of actors/actresses section.

Supplementary Exercise 6.3. Pregnancy rates in cohorts of Ontario girls eligible and ineligible for the HPV vaccination program[7]

1. At entry, what was the age of the youngest girl? the oldest? in what months were these girls born? At the beginning of Grade 10, what was the age of the youngest girl? the oldest?

2. The Lexis diagram in Figure 2 at the end of thse Notes was drawn using the `polygon` primitive in `R`. Instead, using a few individual representatives of each cohort, and using the `Epi` package in `R`, draw lines in a Lexis diagram to depict the entry and the follow-up of of these individuals. If possible, use a different colour for each cohort.

3. Our objective is to fit a rate (i.e. hazard, or incidence density) of pregnancy that is a function of *age*, *calendar year*, and *eligibility* for the program, i.e. λ[age,year,eligibility] Before seeing the data, suggest some candidate forms for λ[age,year,eligibility] and describe the interpretation of each parameter in the model. How are these parameters linked to the parameters in the models fitted by Smith et al.?

4. Fit your suggested model(s) to the data provided by these authors.[8]

[7] Refer to the article 'Effect of human papillomavirus (HPV) vaccination on clinical indicators of sexual behaviour among adolescent girls: the Ontario Grade 8 HPV Vaccine Cohort Study' by Smith et al. CMAJ 2014. DOI:10.1503 /cmaj.140900. A link to this article can be found under the heading 'Link to Smith and Levesque study of post-HPV-vaccine behaviour of Ontario girls' at the bottom of the resources page for C&H chapter 06. It will take you to a page of material about HPV, and to the CMAJ article by Smith et al.

[8] Because the authors were busy preparing grant applications, and because their agreement with the Ontarion data-holders then lapsed, they were unable to provide the 'real' data. So you will have to do with the 'made up' dataset in the 'dummy' table: you should be able to make an 'eye-estimate' of the model parameters without having to enter the data and fit the model via software.

Hi Leah and Linda [email in December 2014]

First, Leah, congratulations on an important study, on the publication, and on the PhD itself! I missed your defense. But I heard about it from others and from Eduardo F. – who I expect was an examiner.

It so happened I spoke in his unit a week or so later, and he mentioned the item about the month of birth.

I have used the Gardasil trials extensively in exercises in my biostat for biostatisticians (601) course, which I model on the Clayton and Hills Statistical Models for Epidemiology book. Since their exercises are very small and a bit too sanitized, above are 2 of the more real-life exercises I work into the material. I also used the adult circumcision studies in Africa, and the Salk polio trial, so they get to hear a good bit about viruses.

I focused on efficacy and protection, whereas you are studying the *downsides*. There is the same worry with adult circumcision that those who become circumcised may undertake more risky behaviours and undo some of the good.

The *pregnancy rates (as hazard functions) would make an ideal application of rate models that have 2 time dimensions*, age and year, and the data are nicely visualized in a Lexis diagram.

Above I have sketched some exercise questions based on your article, and below I show the type of diagram I would expect as an answer.

So I am wondering if you would be willing to 'contribute' some data for this teaching objective. I would not make them publicly accessible (they would just be for teaching), and I would be happy with just counts – no need for individual data, just numbers of outcomes (either STD's or pregnancies or combined) as cell counts, in age-time bins... with bins say 1 month or 3 months wide. obviously 1-month bins would make for interesting modelling of rates as almost continuous functions of both age and year.

I am putting a dummy (but I hope not too unrealistic) table of the counts in the next page, so you can see the type of 'age by calendar year (or school month!) layout from which the counts in the Lexis diagram can be reassembled.. any such array would do, so you could go with what would be easiest for you. Of course, if you would be willing to split the counts across 48 months rather than 16 quarter-years, that would be ideal.

A couple of questions about the follow-up...

You say you counted events from Sept 1 grade **10** to March 31 grade 12. Was there a reason not to count events all the way from Sept grade **8** onwards? i.e. for 4 years and 7 months instead of 2 years and 7 months you used? I know there would not be as many events at these younger ages of follow-up but it would be nice to see how the 'pregnancy rate' functions 'takes off' as a function of age, starting from age 12.67! and going all the way to 18.25 years (you can see why *my focus is age, the dominant factor here*. It would make quite an interesting intro to real epi (and modelling) example for our biostat students...

I was surprised at how high the 2 years and 7 months 'risk' of pregnancy was.. 4% seems a lot, but I expect many are terminated, but still. Is there a strong SES gradient?

Lastly, what attracted my attention initially was Eduardo's remark that you had found the relation with month of birth and that it was puzzling him, and that people were going to look into it further. I expect that you were well aware of the reason for this, since you used the different Jan-Dec cohorts as the grades, and the January ones would always be the oldest in their class and the December ones the youngest.

I had been sensitized to this (so called month of birth) pattern from a few decades back, and in 607 I used to give the attached examples from tennis and european football, and more recently (prompted by the opening story in Malcolm Gladwell's book about the month of birth of) NHL players.

But yours is a new 'classic', and its even closer in interest and age to many of our students than tennis or the NHL. So it will be a great teaching example and a nice teaser for general audiences (we used several of these in the minimed lecture series for the public this Fall, and this would be great one!). And of course, every time I would mention it, I would refer to Smith and Levesque rather than Gladwell!

cf http://www.medicine.mcgill.ca/epidemiology/hanley/minimed/ – MYSTERIES

I did speak about this in a TV interview last year, after Erica and Amee Manges' piece on month of birth and a child's chances of getting into the McGill daycare!

http://globalnews.ca/video/762406/how-your-birth-month-affects-your-future

In Appendix 5, do you mean '2 years **before** cohort entry'?

Anyway, congratulations again on great work. Many people would have believed this result even if you had not done the study, but then there is a big sub-population that would not. Null results are important, and this nailed it!

I expect you will have some restrictions on what data you can share, but I am hoping that pregnancy counts in 31 month of age × 48 month of birth = 1,488 cells will be ok.. It's more detail than the 8 counts I can derive from your paper, but such higher-resolution age-and-year specific counts won't breach privacy. **Best.... Jim**

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]	[,15]	[,16]
[1,]	3	3	4	6	3	4	9	1	3	2	3	0	2	4	3	2
[2,]	5	3	3	6	5	5	3	5	4	3	2	2	6	3	1	3
[3,]	5	1	3	7	5	7	4	7	3	7	3	6	4	1	3	1
[4,]	3	4	4	6	5	8	4	6	1	5	4	2	4	6	3	1
[5,]	3	10	2	8	7	5	7	3	3	7	4	4	4	4	4	4
[6,]	14	8	8	4	5	9	12	6	5	5	6	5	5	3	8	4
[7,]	7	8	4	5	8	10	4	2	7	4	14	9	6	4	5	4
[8,]	7	13	7	4	7	5	6	6	8	7	6	4	4	7	6	4
[9,]	14	6	8	6	17	13	4	9	8	5	9	9	11	6	8	10
[10,]	9	10	10	11	7	11	12	6	8	11	6	5	8	7	3	8
[11,]	13	9	8	12	12	9	9	11	10	6	7	8	9	11	3	11
[12,]	16	10	16	13	16	11	9	5	16	13	4	11	6	8	10	6
[13,]	15	11	7	7	10	18	18	9	12	12	12	10	11	5	10	11
[14,]	14	12	15	18	12	16	9	10	15	11	6	9	11	9	9	7
[15,]	17	19	11	14	17	14	9	10	13	11	12	7	12	9	13	6
[16,]	20	18	14	12	12	20	11	9	10	15	10	11	17	24	21	11
[17,]	14	16	8	23	20	19	14	11	11	14	17	23	18	20	14	11
[18,]	20	19	24	14	20	17	16	11	17	23	16	13	19	20	13	15
[19,]	15	30	22	19	28	21	29	18	20	35	23	16	19	18	18	14
[20,]	22	25	22	18	24	20	23	18	23	30	19	23	14	21	15	16
[21,]	30	22	31	21	30	28	32	28	29	29	29	23	19	22	21	18
[22,]	31	33	39	24	35	24	22	18	22	25	30	18	21	16	16	20
[23,]	38	38	29	21	41	29	35	24	31	31	34	19	32	29	28	22
[24,]	34	43	33	28	39	42	19	33	25	33	33	22	31	33	23	28
[25,]	48	35	38	32	31	27	28	30	28	38	36	28	24	31	28	23
[26,]	46	62	50	31	40	33	42	39	37	28	39	25	33	40	36	25
[27,]	47	59	49	38	52	48	42	36	50	43	37	33	34	38	26	29
[28,]	66	54	41	57	63	55	47	33	50	35	57	38	43	25	49	38
[29,]	59	58	53	45	51	58	37	46	52	41	55	48	35	44	52	39
[30,]	68	70	57	53	65	57	57	49	64	64	55	51	52	55	44	36
[31,]	74	74	65	58	88	71	46	61	74	67	47	45	74	61	51	31

rows [1,] to [31,] are the 31 months from Sept Grade 10 to March Grade 12
columns [,1] to [,16] correspond to the 16 'columns' (3-month cohorts) in Figure 3
(i.e., each is a 3-month bin of births)
Entries are numbers of pregnancies (approximated to match Figure 3B), with a rate that increases
each month within the column by x percent per month -- so the rate in March of Grade 12 when
girls are 30 months older than in Sept of grade 10, is approx. xx.x times higher] (1+x/100) to power of 30].

Counts in 31 rows (as is) and 48 columns (instead of 16) would be ideal!

Inference from 2 way Tables M&M §9 *Test for trend in (Response) Proportions [from A&B §12.2]*

Example **Birth date and sporting success**

SCIENTIFIC CORRESPONDENCE in NATURE • VOL 368 • 14 APRIL 1994 p592

Sir — I have found a significant relationship between birth date and success in tennis and soccer. In the Netherlands and England, players born early in the competition year are more likely to participate in national soccer leagues. The high incidence of elite athletes born in the first quarter of the competition year can be explained by the effects of age-group position.

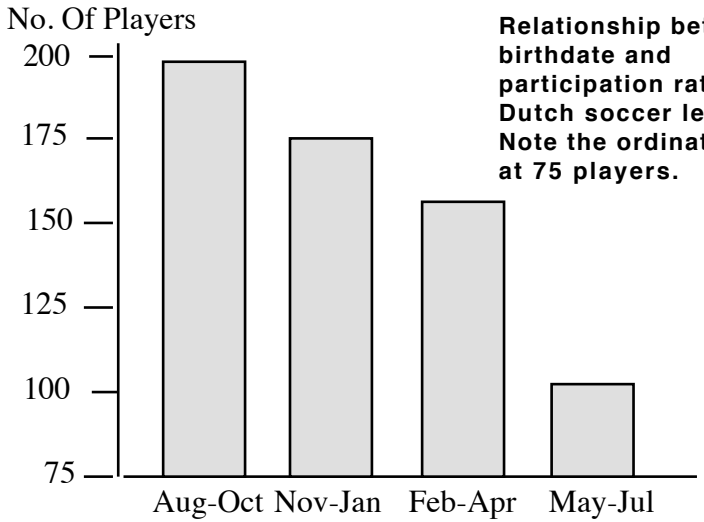
In organized sport. talent is considered predominantly in terms of physical skills. and the influence of social and psychological factors is often ignored or underestimated[1]. Various studies have investigated the psychological characteristics of elite athletes[2], but none has looked for an effect of age. I discovered a strikingly skewed distribution of the dates of birth of 12- to l6-year-old tennis players in the top rankings of the Dutch youth league. Half of a sample of 60 tennis players were born in the first 3 months of the year.

This discovery led me to consider the distribution of the dates of birth of professional soccer players. In the Netherlands, there are two leagues comprising a total of 36 clubs. I found a striking difference between participation rates of those born in August and July. The Dutch soccer competition year starts on the first of August. A chi-square test indicates that the distribution is not uniform (P<0.001); and a regression analysis demonstrates a clear linear relationship between month of birth and number of participants. The dates of birth of 621 players, compiled into quarters, are shown in the figure. This relationship cannot be attributed to the distribution of births in the Netherlands, as this is highly uniform.

We also inspected the distribution of the dates of birth of English football players in league clubs in the period 1991-92 (ref.3). Birth dates for all players were tabulated by month and compiled into quarters. The results (table) show the significant effect of date of birth on participation rate of soccer players within each of the national leagues, indicating that. as in the Netherlands, significantly more football players are born in the first quarter of the competition year (which starts in September in England).

There is a known relationship between date of birth and educational achievement[5]. implying that the younger children in any school year group are at a disadvantage compared to the older children. Children who participate in sports are also placed in age groups, and my results imply many athletes in organized sports may never get a fair chance because of this method of classification. Very little attention has been drawn to this problem. One of the few studies done in this area analysed the dates of birth of young Canadian hockey players in the 1983-84 season[6]. Players possessing a relative age advantage (born in the months lanuary-June) were more likely to participate in minor hockey and more likely to play for top teams than players in July-December.

More than 20 years ago, this journal published an article concerning the relationship between season of birth and cognitive development[7]. The authors attributed this relationship to a fault in the British educational system. A similar relationship was found[5] in the Netherlands. Despite this, no action was undertaken to change the educational system. One can only hope that this will not he the case for sports.



Relationship between birthdate and participation rates in Dutch soccer league. Note the ordinate begins at 75 players.

PARTICIPATION RATES IN ENGLISH SOCCER LEAGUES

	Players in birthdate quarters				Statistics		
League	Sep-Nov	Dec-Feb	Mar-May	Jun-Aug	Total	Chi-Square	Sig. Level
FA premier	288	190	147	136	761	75.5	P<0.0001
Division 1	264	169	154	147	734	48.47	P<0.0001
Division 2	251	168	123	131	673	61.11	P<0.0001
Division 3	217	169	121	102	609	52.38	P<0.0001
Total	1,020	696	545	516	2,777	230.77	P<0.0001

References: **1** Dudink A Fur J High Ability 1, 144-150 (1990). **2** Dudink A & Bakker. F. Ned. Tschr. Psychol 48. 55 -69 (1993). **3** Rollin,J *Rothmans Football Yearbook 1992-93* (Headline. London. 1992). **4** Shearer.E Educ Res 10. 51-56 (1967) **5** Doornbos, K. [Date of birth and scholastic performance (Wolters-Noordhoff, Groningen. 1971). **6** Barnsley. R. H. & Thompson A. H. Can. J. Behav. Sci 20. 167-176 (1988). **7** Williams. Ph.. Davies P., Evans, R & Ferguson, N. Nature 228. 1033-1036 (1970).

Ad Dudink Faculty of Psychology, University of Amsterdam, 1018 WB 3 Amsterdam, The Netherlands

——

For an example of an analysis of seasonal variation, see the article by H T Sørensen et al. Does month of birth affect risk of Crohn's disease in childhood and adolescence? p 907 BMJ VOLUME 323 20 OCTOBER 2001 bmj.com (copy of article, and associated dataset, on course 626 website).

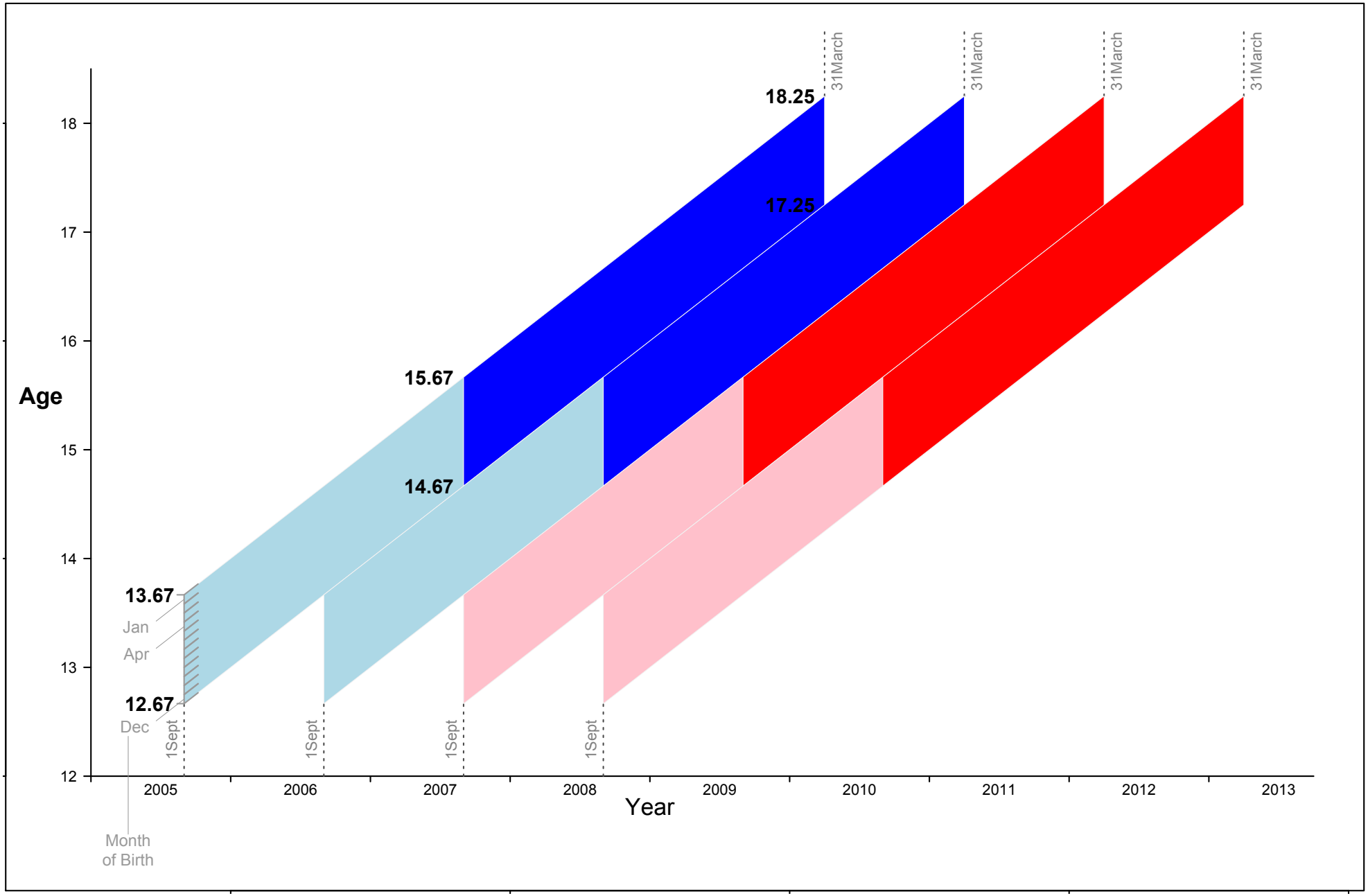


Figure 2: Lexis Diagram, 4 Ontario cohorts. Each cohort is further split into 12-subcohorts according to month of birth. The oldest/youngest in any school grade are those born in January/December.