Practical exercises: Measures of Disease Occurrence Analysis of Epidemiological Data

Nordic Summer School of Cancer Epidemiology
Danish Cancer Society, 12–23 August, 2024
http://BendixCarstensen.com/NSCE/2022
Version 1.0

Compiled Friday 9th August, 2024, 14:41
from: C:\Bendix\teach\NSCE\2024\pracs/pracs.tex

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

Esa Läärä Department of Mathematical Sciences
University of Oulu, Finland
Esa.Laara@oulu.fi

Chapter 1

Introduction to exercises

The exercises in this course requires you to do calculations which in principle can be done on a hand-calculator.

However we assume that you use your laptop and use R as a calculator. This will enable you to take the solutions with you home in the form of a file with computer code that does the analyses. It will also enable you to do analyses repeatedly on slightly different sets of data.

At the end of the course you will get a complete set of solution suggestions. Many of these will be quite elaborate, merely as an illustration of how to use the actually existing features in R to produce solutions. They should not be taken as indications of what we assume that you should be able to do.

So here is an indication of how you should use R:

1.1 What is R?

R is free program for data analysis and graphics. It contains all state of the art statistical methods, and has become the preferred analysis tool for most professional statisticians in the world. It can be used as simple calculator and as a very specialized statistical analysis and reporting machinery.

The special thing about R is that you enter commands from the keyboard into a console window, where you also see the results. This is an advantage because you end up with a script that you can use to *reproduce* your analyses—a requirement in any scientific endeavour.

The disadvantage is that you somehow have to find out what to type. The practicals will contain some hints, and you will mostly be using R as a calculator — type an expression, hit the return key and you get the result on your screen.

1.2 Getting R

You can obtain R, which is free, from CRAN (the **C**omprehensive **R** **A**rchive **N**etwork), at http://cran.r-project.org/. Under "`Download and Install R`" click on "`Download R for Windows`" and then click on "`base`" and further "`Download R 3.4.1 for Windows`", which is a self-extracting installer. This means that if you save it to your computer somewhere and click on it, it will install R for you.

Apart from what you have downloaded there are several thousand add-on packages to R dealing with all sorts of problems from ecology to fiance and incidentally, epidemiology. You

must download these manually. In this course we shall only need the `Epi` package.

1.2.1 Starting R

You start R by clicking on the icon that the installer has put on your desktop. You should edit the properties of this, so that R starts in the folder that you have created on your computer for this course: Right-click on the R-icon, choose "Properties", and then in the field "Start in", enter the relevant folder-name.

Once you have installed R, start it, and in the menu bar click on Packages→Install package(s)..., chose a mirror (this is just a server where you can get the stuff), and the the `Epi` package.

Once R (hopefully) has told you that it has been installed, you can type:

```
library( Epi )
```

to get access to the `Epi` package. You can get an overview of the functions and data sets in the package by typing:

```
library( help=Epi )
```

1.2.2 Quitting R

Type `q()` in the console, and answer "No" when asked whether you want to save workspace image.

1.3 Working with a script editor

1.3.1 Built-in editor in R

If you click on `File→New script`, R will open a window for you which is a text-editor very much like Notepad.

If you write a commands in it you can transfer then to the R console and have them executed by pressing `CTRL-r`. If nothing is highlighted, the line where the cursor is will be transmitted to the console and the cursor will move to the next line. If a part of the screen is highlighted the highlighted part will be transmitted to the console.

1.3.2 Rstudio

is a front-end to Rwith many facilities. It is a commercial product but there is a free version which works excellent with many handy facilities; if you go to their website, https://www.rstudio.com/, it is easy to download and install.

It is becoming the de-facto interface to R so it is a good idea to use it; you will find that it is quite easy to get help on.

1.3.3 Try!

Now open a script by `File→New script`, and type:

```
5+7
pi
1:10
N <- c(27,33,81)
N
```

Run the lines one at a time by pressing `CTRL-r` (if you are using RStudio it is `CTRL-Enter`), and see what happens.

You can also type the commands in the console directly. But then you will not have a record of what you have done. Well, you can press `File`→`Save History` and save all you typed in the console (including the 73.6% commands with errors).

1.4 Getting a bit more training

If you are interested in using R in epidemiology, there is "A short introduction to R", originally written for the European Educational Programme in Epidemiology (and for the IARC summer school in time trends in 2007). A revised version is at:
http://bendixcarstensen.com/Epi/R-intro.pdf.

1.5 Further reading

On the CRAN web-site the last menu-entry on the left is "Contributed" and will take you to a very long list of various introductions to R, including manuals in esoteric languages such as Danish, Finnish and Hungarian.

A very short (12 pages) and handy introduction found there is "A (very) short Introduction to R" by Paul Torfs and Claudia Brauer
https://cran.r-project.org/doc/contrib/Torfs+Brauer-Short-R-Intro.pdf. That will take you a long way.

Chapter 2

Measures of Disease Occurrence — Exercises

2.0 Using NORDCAN

2.0.1 Finding and opening NORDCAN

1. Launch your favourite browser.
2. Enter the website of The NORDCAN Project: https://nordcan.iarc.fr/en

2.0.2 Fact sheet on lung cancer

Create a fact sheet for lung cancer in all NORDCAN countries together by appropriate choices from the pertinent menus. For **1 - Populations** choose NORDCAN Countries, and for **2 - Cancer sites** choose Lung; then click **DOWNLOAD FACT SHEET** on the right. Find answers to the following questions:

1. What were the average annual numbers of new cases in men and women during 2017–21?
2. How big were the estimated risks of getting lung cancer by 75 years of age for the two genders?
3. How many men and women died each year from lung cancer during 2017–2021?
4. What were the numbers of men and women living with lung cancer at the end of 2021, and how big were the corresponding proportions of lung cancer patients out of the whole male and female populations, respectively?
5. Compare the trends of age-standardized incidence and mortality rates in men and women. What kind of observations you make?

2.0.3 Incidence of lung cancer

Learn more about the incidence rates of lung cancer among men in the Nordic Countries during 2017-2021. Click **Data visualization** and within box **Incidence/Mortality** click Tables. Among the menus on the left, choose Lung from **Cancer sites (1)** and click on Period

beside **Year** but keep the defaults offered for the other menus as they are: Incidence, Males, Countries, etc.

1. Where was the age-standardized incidence (ASR, according to World Standard Population) highest, where lowest? What were the crude incidence rates in these two populations?

2. Compare Finland and Norway. Can you find any essential difference between them in the crude rates? What about the age-standardized rates with different standard populations? (The explanation for the standardized rates and for possible discrepancies between them and the crude rates will be given later on.)

2.0.4 Mortality from lung cancer

Learn more about the mortality rates of lung cancer among men in the Nordic Countries during 2017-2021. Proceed as with the incidence of lung cancer above but change the **Measures** into Mortality and execute.

1. Where was the age-standardized (World) mortality highest, where lowest? What were the crude rates in these two populations? Are the standardized and crude rates very different from the corresponding incidence rates above?

2. Compare Island and Sweden. Can you find any real difference in the crude rates? What about the age-standardized rates with different standard populations?

2.0.5 Prevalence of lung cancer

Learn more about the prevalence of lung cancer among men in the Nordic Countries at the end of 2021. Under **Data visualization** go to box **Prevalence**, and continue to Prevalence - Tables. Perform similar choices for the different menus as you did above for incidence and mortality.

1. Where was the total prevalence highest, where lowest? What were the crude prevalence proportions in these two regions?

2. What was the crude prevalence proportion of cases diagnosed less than 5 years ago in all Nordic countries jointly?

3. Can you find out, what were the prevalence proportions of cases diagnosed at least 5 years ago in these populations?

2.0.6 Crude and standardized rates: stomach cancer

Go back to **Data visualization** and obtain the crude and standardized incidence rates of male stomach cancer in the Nordic countries for 2017-2021.

1. In which population is the incidence highest when measured both by the crude rate and by all the different age-standardized rates?

2. Compare the age-standadized rate based on the World Standard Population of the populations in item (a) with those of Cali and Birmingham in the 1980s given on lecture slides, page 75.

3. Why are the standardized rates of type ASR(N) not very different from the crude rates? Why are the ASR(W) and ASR(E) lower when compared to ASR(N)?

2.0.7 Cumulative risk by 75 y: stomach cancer

From the table created above, examine the estimated cumulative risks of male stomach cancer by 75 years of age in each of the Nordic countries for 2017-2021. Look at the last column on the right.

1. Where does this measure seem to be highest and where lowest, and how big these "risks" are are?

2. Compare the figures of these countries with those of Cali and Birmingham during 1980s given on page 75 of lecture slides.

2.0.8 Relative survival

Now we shall have a look at the prognosis of lung cancer patients when compared with the general population. Under **Data visualization** proceed to box **Survival** and click Survival - Tables. A table is opened displaying age-standardized 1-year relative survival (%) [95% CI] for males for all Nordic countries and cancer sites and cases diagnosed during 2017-2021.

1. To compare 5-year relative survival of lung cancer in men between the Nordic countries over time we shall proceed as follows. From menus on the left, change **Display by** to **Period/Country** and within **Cancer sites (1)** choose Lung. Then, on top left of the table, change 1-years to 5-years. – In which country was the relative survival poorest and where was it most favourable among male patients diagnosed during 2012–2016?

2. Now take a look at the 5-years relative survival of lung cancer in women in the Nordic countries diagnosed in 2012-2016 by changing **Sexes** to Females. What is your general observation on the direction of the difference between men and women in each country?

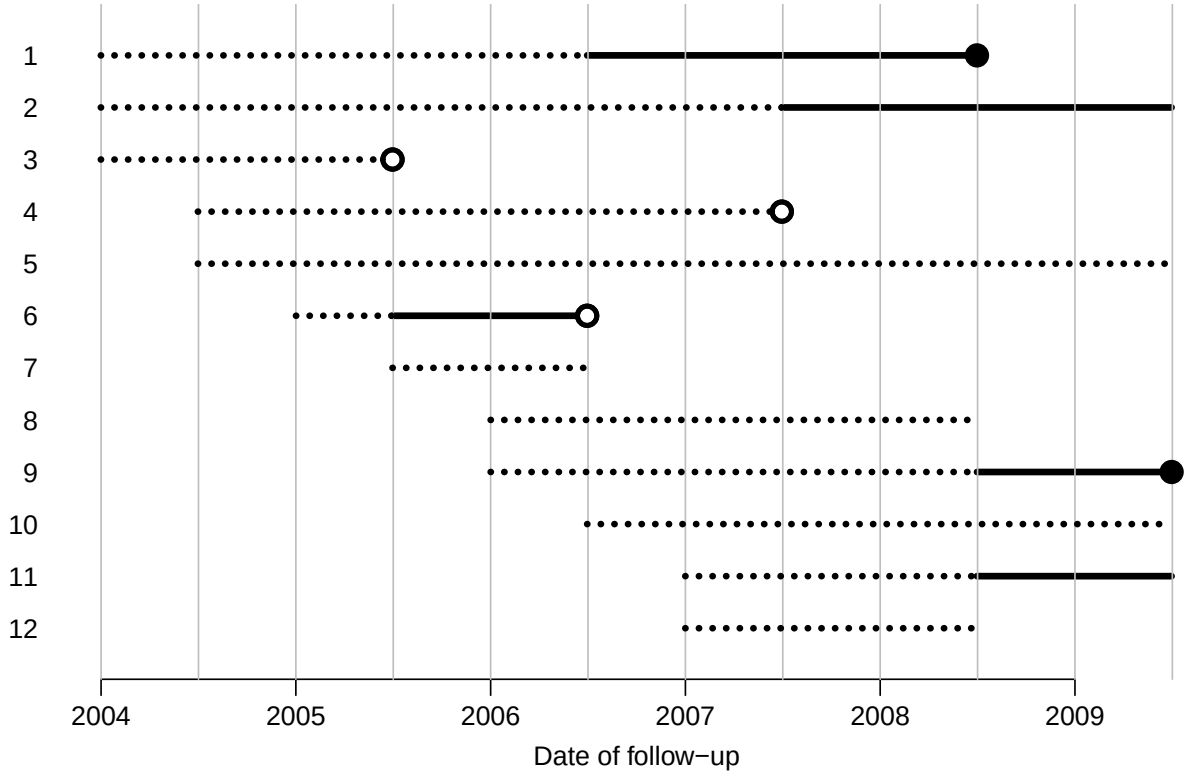
3. Can you figure out, by how many percentage points did the relative survival proportion improve in male patients of Norway during the four decades from 1982-1986 to 2012-2016?

4. The answer to the previous question can be directly obtained by first clicking Survival - Improvements within box **Survival** and after that choosing again Lung plus the pertinent periods from menu **Periods**. – Was this result compatible with what you found out in the previous item?

2.1 Basic measures in a cohort

The figure below shows the follow-up experience of members of a small study cohort between 1 January 2004 to 30 June 2009 from entry to follow-up until death (• if due to cancer C, ∘

for other causes) or censoring (end of line). Follow-up until the occurrence of cancer C is shown with a broken line. For those subjects contracting cancer C, follow-up after diagnosis is shown with a solid line.



We shall calculate the values of the incidence rate of the disease and of various mortality measures

1. What is the incidence rate (per 100 y) of cancer C during the period from 1 Jan 2004 to 31 Dec 2008? – Organize the computations as follows:

   (i) Find out from the figure, what are the individual contributions (in years) of persons 1, 4, 5, and 12 to the total amount of person-time of follow-up pertinent to this task.

   (ii) The total person-time is 27 years. Assign this to variable `Y.todis` writing and running the following command line:
   ```
   > Y.todis <- 27
   ```

   (iii) What is the total number of new cases of cancer C? – Assign this number to variable `Cases` in the same way.

   (iv) Obtain the incidence rate of cancer C assigning its value into variable `Irate` and printing it as follows:
   ```
   > Irate <- 100*Cases/Y.todis
   > Irate
   ```

2. What is the mortality rate from cancer C during the same period? – Proceed with similar steps as above:

(i)-(ii) What is the total person time now? Is it the same as before, or more, or less? Assign this to variable `Y.todth` and run the command.

(iii) What is the total number of deaths from disease C? Assign this to variable `Dth.C`.

(iv) Assign the mortality rate from C into variable `Mrate.C` and print

3. What is the mortality rate from all causes during the same period? Assign the total number of deaths into `Dth.all` and compute the total mortality rate `Mrate.all` applying the same principle as above.

4. What is the estimated 3-year mortality proportion ("risk" of death for a risk period of 3 years since entry) from all causes based on the result in the previous item and assuming the constant rate model? – Apply the following command:
`> Mprop3.all <- 1 - exp( - (Mrate.all/100)*3 )`
and print the result. – Why division by 100 is necessary here?

5. What is the mortality rate `Mrate.pts` during the same period from all causes *among the patients with cancer* C after the onset of C? The person-years for this task can be obtained *e.g.* as follows:
`> Y.distodth <- Y.todth - Y.todis;`
Explain why. Count the pertinent number of deaths, compute the rate and print.

(f) What is the estimated 3-year mortality proportion `Mprop3.pts` after the onset of C among the patients with C?

(g) What is the prevalence of C on 30 September 2006, and on 31 December 2008? – Find out the sizes of the populations `N1` and `N2` as well as the numbers of prevalent cases `C1 and C2` at the two time points, and compute the corresponding prevalence proportions `P1` and `P2`. from these.

Why the incidence or mortality proportions for 3-year or any other risk period, calculated by the simple formula presented on slides 16 and 17, would be problematic in tasks 1 and 2?

Difficult: The follow-up of the cohort is an example of a *multistate* set-up where a person can be in each of 4 possible states: "Alive and well", "Alive with cancer", "Dead from cancer" and "Dead from other causes".

1. Draw four boxes, one for each state, and indicate with arrows the possible transitions between them.

2. Indicate for each arrow how many transitions there were in the cohort.

3. Indicate in the boxes, how many person-years was lived in each box.

4. Identify the calculation of rates in this diagram.

2.2 Population life table

Consider the lifetable for the Danish population for the years 1991–95, in table 2.1.

The survival function in the table can be thought of as number of a hypothetical cohort of 100,000 persons starting at age 0, that will still be alive by age a.

1. Calculate the probability that a 40 year old man reaches age 70 / 80 / 90, respectively.

Table 2.1: *Life table for the Danish population for the period 1991–95.(From: Befolkningens bevægelser 1998, Danmarks Statistik, 2000). $S(a)$: The survival function ($\times 100,000$); $p(a)$: Death probability ($\times 100,000$); $R(a)$: Expected residual life time.*

	Men			Women				Men			Women		
Age	$S(a)$	$p(a)$	$R(a)$	$S(a)$	$p(a)$	$R(a)$	Age	$S(a)$	$p(a)$	$R(a)$	$S(a)$	$p(a)$	$R(a)$
0	100,000	712	72.53	100,000	541	77.84	50	92,470	575	25.72	95,542	400	29.92
1	99,288	59	72.05	99,459	52	77.27	51	91,938	606	24.86	95,159	434	29.03
2	99,230	33	71.09	99,407	32	76.31	52	91,381	642	24.01	94,746	464	28.16
3	99,197	30	70.11	99,375	22	75.33	53	90,795	728	23.16	94,306	506	27.29
4	99,168	26	69.14	99,353	19	74.35	54	90,133	829	22.33	93,829	561	26.42
5	99,142	22	68.15	99,335	15	73.36	55	89,386	909	21.51	93,302	618	25.57
6	99,121	20	67.17	99,319	14	72.37	56	88,573	991	20.70	92,726	683	24.73
7	99,101	23	66.18	99,305	14	71.38	57	87,696	1,136	19.91	92,093	765	23.89
8	99,079	25	65.20	99,291	15	70.39	58	86,700	1,315	19.13	91,388	841	23.07
9	99,055	20	64.21	99,276	14	69.40	59	85,560	1,431	18.38	90,619	940	22.26
10	99,035	18	63.22	99,263	11	68.41	60	84,335	1,595	17.64	89,767	1,052	21.47
11	99,017	17	62.24	99,252	13	67.42	61	82,990	1,804	16.92	88,823	1,132	20.69
12	99,001	20	61.25	99,239	14	66.43	62	81,493	1,924	16.22	87,817	1,215	19.93
13	98,981	24	60.26	99,225	14	65.44	63	79,925	2,070	15.53	86,750	1,326	19.16
14	98,957	26	59.27	99,211	17	64.45	64	78,271	2,290	14.84	85,600	1,461	18.42
15	98,931	36	58.29	99,195	19	63.46	65	76,478	2,494	14.18	84,349	1,596	17.68
16	98,896	49	57.31	99,175	21	62.47	66	74,571	2,780	13.53	83,003	1,711	16.96
17	98,847	61	56.34	99,154	23	61.48	67	72,498	3,045	12.90	81,583	1,848	16.25
18	98,787	76	55.37	99,132	32	60.50	68	70,290	3,336	12.29	80,075	2,015	15.54
19	98,711	95	54.41	99,100	41	59.52	69	67,945	3,752	11.70	78,462	2,187	14.85
20	98,618	93	53.46	99,059	36	58.54	70	65,396	4,058	11.13	76,746	2,361	14.17
21	98,526	87	52.51	99,023	32	57.56	71	62,742	4,420	10.58	74,934	2,621	13.50
22	98,441	90	51.56	98,991	35	56.58	72	59,969	4,864	10.05	72,970	2,873	12.85
23	98,352	87	50.60	98,957	33	55.60	73	57,052	5,291	9.54	70,874	3,078	12.22
24	98,266	91	49.65	98,924	30	54.62	74	54,033	5,778	9.04	68,692	3,316	11.59
25	98,177	102	48.69	98,894	35	53.64	75	50,911	6,271	8.57	66,415	3,676	10.97
26	98,076	106	47.74	98,860	41	52.65	76	47,718	6,783	8.11	63,973	4,074	10.37
27	97,972	105	46.79	98,820	40	51.67	77	44,481	7,346	7.66	61,367	4,370	9.79
28	97,869	112	45.84	98,780	42	50.70	78	41,214	8,030	7.23	58,685	4,818	9.20
29	97,759	119	44.89	98,738	48	49.72	79	37,904	8,710	6.82	55,858	5,365	8.66
30	97,643	125	43.94	98,690	52	48.74	80	34,603	9,471	6.42	52,861	5,925	8.12
31	97,522	134	43.00	98,639	60	47.77	81	31,326	10,389	6.04	49,729	6,610	7.60
32	97,391	150	42.06	98,580	65	46.79	82	28,071	11,293	5.68	46,442	7,451	7.10
33	97,245	159	41.12	98,516	61	45.82	83	24,901	12,149	5.34	42,982	8,337	6.63
34	97,090	158	40.18	98,456	72	44.85	84	21,876	13,043	5.01	39,398	9,230	6.19
35	96,936	168	39.25	98,385	90	43.88	85	19,023	14,200	4.69	35,762	10,137	5.77
36	96,773	187	38.31	98,297	105	42.92	86	16,321	15,642	4.38	32,137	11,407	5.36
37	96,592	210	37.38	98,194	118	41.97	87	13,768	17,076	4.10	28,471	12,688	4.99
38	96,390	228	36.46	98,078	119	41.02	88	11,417	18,402	3.84	24,858	13,835	4.64
39	96,170	251	35.54	97,961	131	40.06	89	9,316	20,246	3.59	21,419	15,391	4.30
40	95,928	283	34.63	97,833	157	39.12	90	7,430	21,659	3.37	18,123	16,864	4.00
41	95,657	296	33.73	97,680	164	38.18	91	5,821	22,775	3.17	15,066	18,541	3.71
42	95,374	293	32.83	97,520	176	37.24	92	4,495	24,923	2.96	12,273	20,439	3.44
43	95,094	304	31.92	97,348	201	36.30	93	3,375	26,578	2.77	9,765	22,521	3.19
44	94,806	323	31.02	97,153	211	35.38	94	2,478	28,725	2.59	7,565	24,601	2.97
45	94,500	347	30.12	96,948	231	34.45	95	1,766	30,641	2.44	5,704	26,453	2.78
46	94,171	383	29.22	96,724	264	33.53	96	1,225	33,252	2.30	4,195	28,752	2.60
47	93,810	431	28.33	96,468	293	32.61	97	818	34,446	2.19	2.989	30,269	2.44
48	93,406	478	27.45	96,186	316	31.71	98	536	33,589	2.08	2,084	31,732	2.29
49	92,959	527	26.58	95,882	355	30.81	99	356	37,944	1.88	1.423	35,125	2.12

The **M**edian **R**esidual **L**ifetime is the time which half of the (currently living part of the population) will survive and the other half not.

2. Find the MRL for men and women aged 40, respectively.

2.3 Incidence and mortality – acute leukaemia

In the table below are given the size (in 1000s) of the male population in Finland aged 0-14 years (the age range of "childhood" in pediatrics!) on the 31 December in each year from 1991 to 2000.

Year	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
Population	493	495	496	497	496	495	491	485	481	478

The following numbers of cases describe the incidence of and mortality from acute leukaemia in this population for two calendar periods: 5 years 1993 to 1997 (source: NORDCAN), and year 1999 only (source: Finnish Cancer Registry http://www.cancerregistry.fi/).

	1993-97	1999
New cases of acute leukaemia	113	26
Deaths from acute leukaemia	22	3

1. Calculate the incidence rates of acute leukaemia in this population for the two periods.
2. Calculate similarly the mortality rates of leukaemia.
3. Is there evidence about any change in the incidence and/or mortality between these two periods?
4. What would you conclude about the fatality of leukemia in children?

2.4 ATCB-trial — prostate cancer

The Alpha Tocopherol Beta Caroten (ATBC) Prevention Trial (*N Engl J Med 1994*; **330**: 1029-35) addressed among other things the possible benefits of daily intake of vitamin E supplements in reducing the incidence of cancer among male smokers. The study population of 29,133 regularly smoking 50-69 years old Finnish men were randomized into two groups: active treatment (vitamin E supplementation), and placebo (no supplementation). The following results were obtained for cancer of the prostate after an average follow-up time of 6 years:

treatment group	number of cases	incidence rate (per 10000 years)
vitamin E supplementation	99	11.6
no supplementation	151	17.8

1. Calculate the person-years at risk in the two study groups separately.

2. Estimate the "relative risk" (using incidence rate ratio) and "excess risk" (using rate difference) for measuring the effect of daily supplementation with vitamin E on the risk prostate cancer.

3. Estimate either the attributable fraction or preventive fraction, whichever more appropriate, to describe the proportional impact of vitamin E supplementation.

4. Discuss the results. What can be concluded from these estimates?

2.5 Comparative measures – smokers vs. non-smokers

In the table below you see the mortality rates (per 1000 person-years, age-adjusted) from three important causes of death among life-long non-smokers and regular smokers as observed after 30 years follow-up of a large occupational cohort (men only).

	lung cancer	other lung diseases	cardiovascular diseases
smokers	2.0	3.0	15.0
non-smokers	0.2	1.0	9.0

1. Calculate for each cause of death the following effect measures for comparison between smokers and non-smokers:

   (a) "excess risk", *i.e.* rate difference,

   (b) "relative risk", *i.e.* rate ratio,

   (c) attributable fraction.

2. Discuss the results. What can be inferred about the biological strength and the public health impact, respectively, of regular smoking regarding the three diseases.

2.6 Infant mortality

During 1978 in Finland 269 boys died at the age of <1 year. The size of this male age group was 33,200 on 31 Dec 1977, and on 31 Dec 1978 it was 32,500. The number of boys born alive during 1978 was 32,800.

1. Calculate the mortality rate (per 1000 person-years) in this age group of boys in the year 1978 by the usual method.

2. In national vital statistics the *infant mortality rate* (IMR) is commonly computed as:

$$\text{IMR} = \frac{\text{no. of deaths in age group} < 1 \text{ year during a calendar year}}{\text{no. of live born children during the year}} \times 1000$$

   Calculate the value of this measure for Finnish boys in 1978 from the given data and compare it with the result in item 1.

3. Is the "infant mortality rate" in item 2 indeed a rate as defined in the lectures — why or why not? Is it a proportion?

2.7 Standardization: Colon cancer

Age specific data on the incidence of colon cancer in male and female populations of Finland during 1999 are given in the following table

	Males				Females				
Age group	Cases	Mid-popul. (1000s)	% of all	Rate ($/10^5$y)	Cases	Mid-popul. (1000s)	% of all	Rate ($/10^5$y)	Rate ratio M/F
0–34	10	1157	46.0	**0.9**	22	1109	41.9	**2.0**	0.44
35–54	76	809	32.0	**9.4**	68	786	29.7	**8.6**	1.09
55–74	305	455	18.0	**67**	288	524	19.8	**55**	1.22
75+	201	102	4.0	**196**	354	229	8.6	**155**	1.27
All	592	2523	100		732	2648	100		

Calculate the following summary measures:

1. crude incidence rate in both populations and the rate ratio: males **vs.** females,
2. age-standardized rates and their ratio using the male population as the standard,
3. age-standardized rates and their ratio using the World Standard Population,
4. cumulative rates up to 75 years and their ratio,
5. cumulative risks up to 75 years and their ratio.

Compare and comment the results obtained in items 1 to 3.

Hint: Organize the calculations needed for summary measures such that the necessary age-specific quantities are assigned into pertinent vectors, *e.g.* age-specific rates in women:

```
ratesF.a <- c(2.0, 8.6, 55, 155)
```

and weights from the male population:

```
wM <- c(46, 32, 18, 4)
```

and make use of the `sum()` function of R, for example, when computing the age-standardized rate for women:

```
stdRateF_wM <- sum( wM * ratesF.a ) / sum( wM )
```

2.8 Standardized rates

Below is the number of cases (D) and the age-specific incidence rates (in cases per 100,000 person-years) from the Danish Cancer Register for the period 1983–87 for colon cancer, rectum cancer and lung cancer, by sex.

	Colon				Rectum				Lung			
	Men		Women		Men		Women		Men		Women	
Age	D	Rate	D	Rate	D	Rate	D	Rate	D	Rate	D	Rate
0- 4	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
5- 9	2	0.25	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
10-14	0	0.00	1	0.11	0	0.00	0	0.00	0	0.00	0	0.00
15-19	3	0.30	7	0.73	1	0.10	0	0.00	1	0.10	0	0.00
20-24	4	0.39	8	0.82	1	0.10	1	0.10	8	0.78	4	0.41
25-29	13	1.36	5	0.55	2	0.21	3	0.33	4	0.42	1	0.11
30-34	18	1.89	27	2.96	11	1.15	4	0.44	7	0.73	14	1.53
35-39	50	4.81	38	3.83	19	1.83	26	2.62	46	4.43	35	3.52
40-44	51	5.42	75	8.29	43	4.57	29	3.21	116	12.32	109	12.05
45-49	94	12.68	124	16.92	81	10.92	75	10.24	262	35.33	209	28.52
50-54	173	26.23	231	34.36	157	23.81	104	15.47	592	89.76	421	62.62
55-59	316	49.31	338	50.22	273	42.60	193	28.67	1089	169.95	650	96.57
60-64	492	78.05	511	73.67	402	63.77	251	36.19	1884	298.86	795	114.62
65-69	737	134.35	695	109.04	533	97.16	369	57.89	2206	402.13	843	132.26
70-74	870	189.61	1006	171.59	601	130.99	430	73.34	2308	503.02	773	131.85
75-79	853	267.27	1081	225.24	539	168.88	427	88.97	1824	571.51	621	129.39
80-84	602	342.50	903	281.20	312	177.51	318	99.03	891	506.93	336	104.63
85-89	279	359.19	522	316.19	180	231.73	184	111.45	305	392.66	135	81.77
90+	95	347.54	174	263.40	67	245.11	79	119.59	62	226.82	40	60.55

The effective population size in the period is 2,521,177 men and 2,596,061 women.

The data are available as the file `std-rates.txt` in the course folder; you can read it into R using:

```
> std <- read.table( "std-rates.txt", header=T )
```

1. How many person-years was accumulated by the Danish men aged 70–79 in the period 1983–87 ?
2. Calculate the crude rates for each sex and site.
3. Calculate the cumulative rates to ages 65, 70, 75 and 80.
4. Calculate the standardized rates, standardized to the world standard population:

Age	Weight (×1000)	Age	Weight (×1000)	Age	Weight (×1000)
0– 4	120	35–39	60	70–74	20
5– 9	100	40–44	60	75–79	10
10–14	90	45–49	60	80–84	5
15–19	90	50–54	50	85–89	3
20–24	80	55–59	40	90+	2
25–29	80	60–64	40		
30–34	60	65–69	30		

5. Calculate the male-female ratios of the crude, the standardized and the cumulative rates. Why are they not the same?
6. Calculate the age-specific male-female rate-ratios. Comment on the results.

2.9 Survival: cancer of the tongue

The survival of males in Finland with cancer of the tongue diagnosed during 1967-74 was studied by Hakulinen *et al.* (1981). Sizes of risk sets, numbers of deaths and losses (censorings) tabulated into 1 year subintervals since the diagnosis are given in the following table.

Year of FU	size of risk set	no. of deaths	no. of losses	effect. denom.	prop. deaths	prop. surviv.	cumul. survival
0–	130	45	7				0.644
1–	78	24	9	73.5		0.673	
2–	45	5	7	41.5			0.382
3–	33	2	6		0.067		
4–	25	1	5				
5–	19	–	7	15.5	0.0	1.0	0.340
6–	12	–	6				

1. Complete this table by appropriate figures using the actuarial life table method.
2. Based on the results obtained above draw a survival curve and estimate graphically the median and the quartiles, if possible, of the survival time distribution.

2.10 Conditional survival

For Danish patients diagnosed with cancer of colon and rectum in the period 1978–87 we found the following probabilities of death (in %):

Year from	Colon		Rectum	
diagnosis	Men	Women	Men	Women
1st	43.44	42.13	36.60	34.29
2nd	22.80	19.11	24.00	21.86
3rd	16.74	14.60	21.02	15.67
4th	13.84	10.62	15.59	13.54
5th	11.00	8.69	14.55	11.40
6th	10.13	7.36	9.95	11.17
7th	8.67	5.65	11.37	8.99
8th	7.97	5.51	8.69	8.55
9th	7.42	5.37	10.07	8.14
10th	7.75	5.94	5.16	7.26
11th	4.91	5.66	7.14	2.57
12th	6.72	5.42	6.06	5.63
13th	6.20	6.25	5.00	2.13

1. Calculate for each of the groups the cumulative probability of surviving 1, 3, and 5 years respectively.

2. Calculate the *conditional* probabilities of surviving 3 and 5 years after diagnosis *given* that a Danish patient already has survived 1 year.

From Young, Ries & Pollack: "Cancer Patient Survival Among Ethnic Groups in the United States", JNCI, vol 73, pp. 341–52, we find that for white anglosaxons the cumulative survival probabilities for colon and rectum cancer patients diagnosed 1973–79 in the SEER areas are (in %):

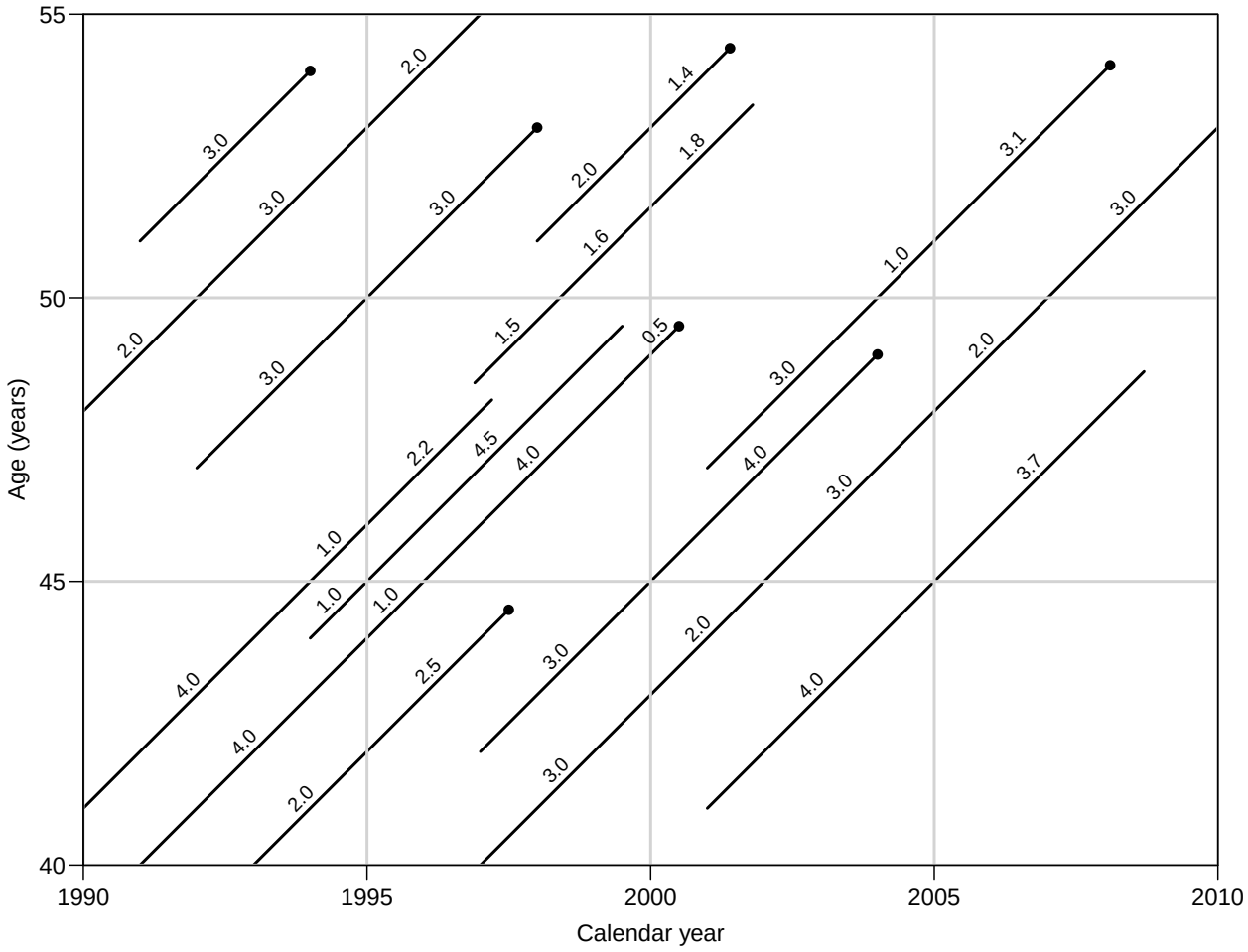
Years from	Colon		Rectum	
diagnosis	Men	Women	Men	Women
1	68	69	74	74
3	46	48	48	50
5	36	39	35	39

3. Calculate the *conditional* probabilities of surviving 3 and 5 years after diagnosis *given* that a U.S. patient already has survived 1 year.

4. Compare the cumulative survival probabilities and the conditional survival probabilities given survival of the first year between Denmark and USA.

2.11 Lexis diagram

In the Lexis diagram below displayed follow-up times of a small occupational cohort over the years 1990-2009 and the age range 40-54 years (this example is modified from a similar one in **B&D**). Each line runs from the entry to follow-up until either the diagnosis of cancer (•), or censoring or withdrawal (no symbol) due to death from other causes or migration.



1. Calculate the numbers of new cases of cancer, and person-years at risk in all the three 5-year agebands: 40-44, 45-49, and 50-54 years for each of the 5-year calendar periods 1990-94, 1995-99, and 2000-04 separately.

   Hint 1: Execute some division of labour in your group, so that not everybody is calculating these items for all periods.

   Hint 2: The data set is available as an example dataset, `occup`, in the `Epi` package. Try:

```
> library( Epi )
> ### data( occup )
> occup <- read.table("http://BendixCarstensen.com/NSCE/R/occup.txt", header=TRUE)
> str( occup )
> occup
> ### example( occup )
```

2. Calculate the numbers of new cases of cancer, person-years at risk in the three 5-year age groups: 40-44, 45-49, and 50-54 years for a *birth cohort* born in 1952-61.

3. Continuing from the previous item, estimate the cumulative rate and the cumulative risk over the whole 15-year age range for the chosen birth cohort.

NB. Estimation of the cumulative risk by the simple formula, presented on lecture slide 63, in which the competing risk of death is ignored, is not so problematic here, because of the relatively young age range covered, in which the mortality is expected to be quite low.

4. The age-specific incidences (per 100,000 person-years) in the three 5-year age-groups during 1990–2010 in the whole population of the country were 100, 200, and 400, respectively, so there was no variation between the subperiods. Assuming that this is an appropriate reference population, calculate the expected number of cases for the index occupational cohort for the same period. Compare the observed and expected number of cases by standardised incidence ratio, SIR.

   Comment on the result.

2.12 Cumulative rates

In the period 1935–47 a large number of persons undergoing cerebral angiography were injected with Thorotrast, a contrast medium containing radioactive Thorium. In order to assess the elevation of the mortality related to the injection of Thorotrast, a control group of patients was selected who had also undergone cerebral angiography on similar indications in the period 1946–63, but with another contrast medium.

Below is a table of deaths and person-years at risk for the two groups, by current age.

	Thorotrast		Controls	
Current age	No. Deaths	Person-years	No. deaths	Person-years
0–19	5	572.1	11	1536.1
20–29	17	1974.2	16	2449.1
30–39	58	3489.0	35	4228.8
40–49	100	4502.2	67	5822.3
50–59	184	4433.5	137	6647.0
60–69	205	2998.1	211	5780.3
70–79	137	1134.4	206	3113.6
80+	45	261.5	114	939.8
Total	751	19365.4	797	30517.6

Calculate the following three things:

1. The estimates of the overall rates in each of the two groups and the rate ratio.

2. A confidence interval for the rate-ratio between the two groups.

3. The cumulative rates to 70 and 80 years in the two groups.

4. The ratio of the cumulative rates.

5. Comment on the results.

2.13 Attributable risk

Consider again the Thorotrast-study material from exercise 2.12 Remember the definition and interpretation of Attributable risk from the lectures.

1. Calculate the attributable risk of Thorotrast exposure on death of patients undergoing cerebral angiography:

   (a) Based on the crude relative risk.

   (b) Based on the relative risk from the cumulative rates to age 70.

   (c) Based on the relative risk from the cumulative rates to age 80.

   Comment on the differences, and calculate the number of deaths attributable to Thorotrast in the three cases.

2. Calculate the attributable risk in each age-group.

3. Calculate the number of deaths attributable to Thorotrast in each group, and compare the sum to the previous results.

Chapter 3

Analysis of Epidemiological Data — Exercises

3.1 Single incidence rates

In Kuwait during 1987 six deaths from stomach cancer were registered in males aged 45 to 54 years, and 89 000 men of this age group were living in the country at that time. In Egypt the corresponding figures in the same male age group during 1987 were 53 cases and 1 819 000 men. Calculate for both countries the following quantities:

1. mortality rate,
2. 95% confidence interval of the "true" rate based on SE of the rate (and error margin),
3. 95% confidence interval of the rate based on SE of the log-rate (and error factor). Compare this with the interval obtained in 2.

3.2 Non-significant difference

A cohort of electric engineers, graduated from a certain university of technology during a specified time interval, were followed-up over a period of 50 years. One out of the 10 female graduates and 1 out of the 200 male graduates developed breast cancer during the follow-up. The difference in the incidence between males and females was "not statistically significant" ($P > 0.05$).

How should this result be interpreted? Choose one from the following alternatives:

1. The results provide supporting evidence for the hypothesis no real difference between males and females in the breast cancer risk among electric engineers.
2. The results are consistent with the universal observation that the risk of breast cancer among females is clearly higher than that in males.
3. No conclusion can be made from this result concerning the male/female contrast in breast cancer incidence among graduates of electric engineering.
4. Other conclusion, what?

3.3 Preventive trial

Read the following abstract of the ATBC Cancer Prevention Study and Figure 2 in it (here shown as figure 1), displaying its major results on cancer incidence, and do the following tasks:

1. State the study hypothesis and the corresponding null hypothesis concerning the effect of receiving daily beta carotene supplements vs. not receiving them on the incidence of lung cancer.

2. Calculate the person-years in the group receiving ceta carotene supplements (the "exposed") and in the group receiving placebo ("unexposed").

3. Calculate the point estimate and the 95% confidence interval for the hazard rate ratio $\rho = \lambda_1/\lambda_0$ of lung cancer between the exposed and the unexposed.

4. Calculate the point estimate and the 95% confidence interval for the hazard rate difference $\delta = \lambda_1 - \lambda_0$ of lung cancer between the exposed and the unexposed.

5. Calculate a test statistic and the associated P value corresponding to the null hypothesis stated in item (a).

6. Discuss the results. Can the estimated relative rate be confounded by age and/or smoking, as the analysis was not stratified by these factors?

The Effect of Vitamin E and Beta Carotene on the Incidence of Lung Cancer and Other Cancers in Male Smokers

The Alpha-Tocopherol Beta Carotene Cancer Prevention Study Group

Background: Epidemiologic evidence indicates that diets high in carotenoid-rich fruits and vegetables, as well as high serum levels of vitamin E (alpha-tocopherol) and beta carotene, are associated with a reduced risk of lung cancer.

Methods: We performed a randomized, double-blind, placebo-controlled primary-prevention trial to determine whether daily supplementation with alpha-tocopherol, beta carotene, or both would reduce the incidence of lung cancer and other cancers. A total of 29,133 male smokers 50 to 69 years of age from southwestern Finland were randomly assigned to one of four regimens: alpha-tocopherol (50 mg per day) alone, beta carotene (20 mg per day) alone, both alpha-tocopherol and beta carotene, or placebo. Follow-up continued for five to eight years.

Results: Among the 876 new cases of lung cancer diagnosed during the trial, no reduction in incidence was observed among the men who received alpha-tocopherol (change in incidence as compared with those who did not, −2 percent; 95 percent confidence interval, −14 to 12 percent). Unexpectedly, we observed a higher incidence of lung cancer among the men who received beta carotene than among those who did not (change in incidence, 18 percent; 95 percent confidence interval, 3 to 36 percent). We found no evidence of an interaction between alpha-tocopherol and beta carotene with respect to the incidence of lung cancer. Fewer cases of prostate cancer were diagnosed among those who received alpha-tocopherol than among those who did not. Beta carotene had little or no effect on the incidence of cancer other than lung cancer. Alpha-tocopherol had no apparent effect on total mortality, although more deaths from hemorrhagic stroke were observed among the

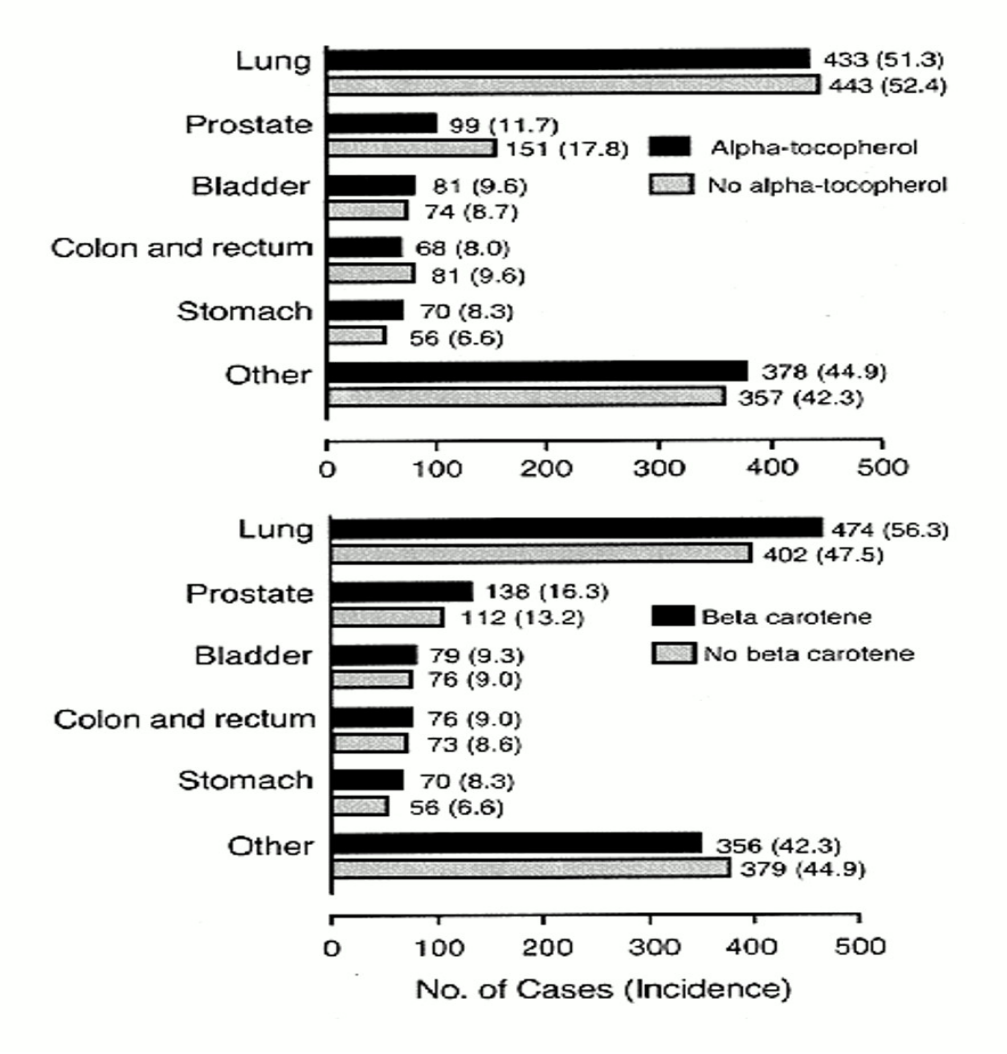


Figure 3.1: *Number and Incidence (per 10 000 Person-Years) of Cancers, According to Site, among Participants Who Received Alpha-Tocopherol Supplements and Those Who Did Not (Upper Panel) and among Participants Who Received Beta Carotene Supplements and Those Who Did Not (Lower Panel).*

men who received this supplement than among those who did not. Total mortality was 8 percent higher (95 percent confidence interval, 1 to 16 percent) among the participants who received beta carotene than among those who did not, primarily because there were more deaths from lung cancer and ischemic heart disease.

Conclusions: We found no reduction in the incidence of lung cancer among male smokers after five to eight years of dietary supplementation with alpha-tocopherol or beta carotene. In fact, this trial raises the possibility that these supplements may actually have harmful as well as beneficial effects.

(*New England Journal of Medicine*, Volume 330, pp. 1029–1035, April 14, 1994, Number 15).

3.4 Preventive trial — interpretation

We continue with the ATBC Cancer Prevention Study complementing its results with those of two other randomized trials that addressed the same hypothesis on the possible beneficial effect of beta caroten supplementation on lung cancer incidence.

1. In the ATBC study the observed rate ratio of lung cancer associated with daily intake of beta caroten supplement appeared to be "statistically significantly" different from 1 ($P = 0.01$). However, the direction of the estimated rate ratio was opposite to that of the original study hypothesis, which was based on the observational evidence that motivated the trial.

   Do you think that this result provides a sufficient basis to conclude that beta caroten supplementation is actually harmful?

2. In the *Beta Carotene and Retinol Efficacy Trial* conducted in USA, a total of 18 314 smokers, former smokers, and workers exposed to asbestos were randomized into two groups: active-treatment group and placebo group (*N Engl J Med* 1996; 334: 1150-1155). The active-treatment group received a combination of 30 mg of beta carotene per day and 25 000 IU of retinol (vitamin A) in the form of retinyl palmitate per day. After a follow-up of 4.0 years on average, the active-treatment group had a relative rate of lung cancer of 1.28 (95 % CI, 1.04 to 1.57; $P = 0.02$) as compared with the placebo group.

   Taken this result together with that of the ATBC trial, what can we now say about the accumulated evidence on the effects of beta caroten on the incidence of lung cancer among smokers? Would we now be more convinced about the harmfulness of this form of vitamin supplementation?

3. A third beta caroten trial was conducted in a study population of 22071 male American physicians (*N Engl J Med* 1996; 334: 1145-1149). After 13 years follow-up the point estimate of the rate ratio of lung cancer between the beta caroten and the placebo groups among the subset of current smokers in that study population was 0.9, *i.e.* lower than 1 but "non-significant" (95% CI 0.58-1.40, $P = 0.63$).

   Is this result in conflict with the results of the two other trials quoted above?

4. In the American physicians' study, among *nonsmokers* the observed rate ratio of lung cancer between beta caroten and placebo groups was 0.78 (95% CI 0.34-1.79, $P = 0.56$).

   What can we conclude about the effect of beta caroten supplementation in non-smoking men on the basis of these results? Is it different from that among regular smokers?

3.5 Geographical variation

Geographical variation in the incidence of certain form of cancer D in a country C was mapped using two classifications for dividing the area: (a) by county, and (b) by central hospital district. In the figure 2 the adjusted incidences (per 100,000 person years) of D are given for certain areas according to both divisions.

In addition are given stars indicating that the figure in question is significantly different ($p < 0.01$) from the average incidence of D in the whole country, which was 1 per 100,000

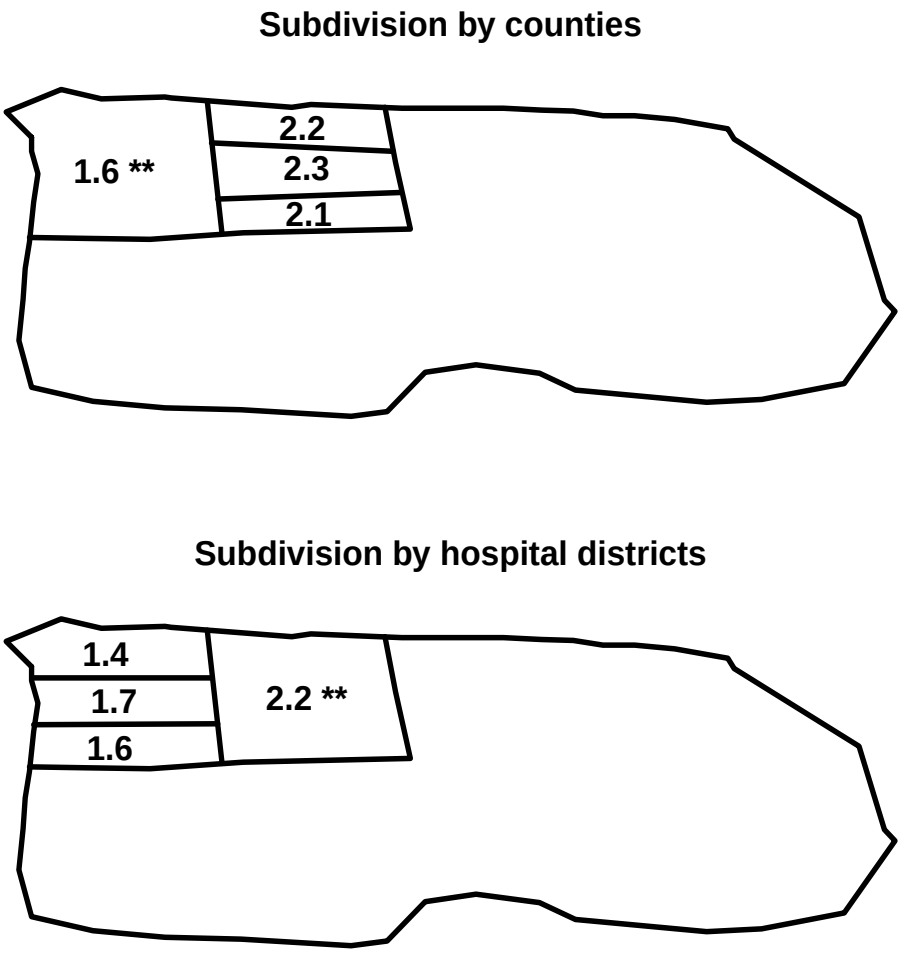


Figure 3.2: *Geographical division by county (top) and hospital district (bottom).*

person-years. The two divisions seem to give somewhat contradictory results. How can we explain this apparent paradox?

3.6 Efficiency of study design

You are designing a cohort study to estimate the relative risk associated with a certain exposure factor X. Initially you are planning to recruit 10 000 persons to the cohort, such that 2000 would be exposed and 8000 unexposed to X, and you intend to have a 5 year follow-up period. A statistician points out that the confidence interval of your relative risk estimate is likely to be too wide. You cannot afford to enroll more than 10 000 individuals to the cohort. How could you change your research plan in principle such that the confidence interval would become shorter without increasing the total number of study subjects?

3.7 Case-control study: MI

In the table below are results presented from an unmatched case-control study on the association between physical activity (PA) and risk of myocardial infarction (MI) stratified by gender.

Table 3.1: *Table of cases and controls by sex and PA (physical activity) index*

Gender	PA index	Cases	Controls	Total
Men	2500+ kcals	141	208	349
	< 2500 kcals	144	112	256
Total		285	320	605
Women	2500+ kcals	49	58	107
	< 2500 kcals	32	45	77
Total		81	103	184
Both	2500+ kcals	190	266	456
	< 2500 kcals	176	157	333
Total		366	423	789

1. Calculate the point estimate (and the 95% confidence interval) of the rate ratio in both genders separately.
2. What can you say of the possible modification of the effect of PA by gender; is the relative risk different in males than in females?
3. Is gender a confounder for the association between PA and MI; on what grounds?
4. Calculate the crude point estimate of the rate ratio, unadjusted for gender.
5. Calculate the gender-adjusted summary estimate of the rate ratio (and its 95 % confidence interval), using `glm` with binomial error as indicated in the lecture slides.
6. Compare this with the crude one.
7. Is there effect-modification by sex?
8. How would you report this?

3.8 Case-control study: Neonates

Cnattingius *et al.* (*JNCI 1995*; 87 (June 21): 908-914) reported a case-control study on prenatal and neonatal risk factors for childhood lymphatic leukaemia in children. From the

National Cancer Register of Sweden they collected all cases of this disease reported in children under 15 years of age from 1973 through 1989. Five controls for each case, matched for age and gender, were obtained from the Medical Birth Register of Sweden. The data on potential risk factors in both cases and controls were obtained from the latter register, too.

One of the findings was that 8 children with leukaemia and 2 of the control children had Down's syndrome.

1. On the basis of this information only, can you obtain any reasonable approximations for the following quantities:

   (a) a crude estimate of the relative hazard of leukemia in children with Down's syndrome as compared with children without this chromosome abnormality,

   (b) an approximate 95% confidence interval for the hazard ratio. What assumptions are needed in order that these approximations would be credible?

2. What additional data would be needed to obtain adequate estimates and confidence intervals?

3.9 Matched case-control study: Chemicals

A certain chemical exposure E was studied as a potential risk factor of cancer D in a case-control study with 20 cases and 20 controls. The following observations were made on the exposure status (+ = exposed, − = nonexposed) of each case and control:

No.	case	control	No.	case	control
1.	+	−	11.	−	+
2.	+	−	12.	+	+
3.	−	−	13.	+	−
4.	+	+	14.	−	−
5.	−	+	15.	+	−
6.	+	−	16.	+	−
7.	+	−	17.	+	−
8.	+	−	18.	+	+
9.	+	+	19.	−	−
10.	−	−	20.	+	−

1. Calculate the point estimate (with the approximate 95% confidence interval) of the hazard rate ratio associated with the exposure, as well as the test statistic and P-value corresponding to the null hypothesis of no effect, assuming that the study subjects have been obtained

   (a) by choosing the control group as a random sample of the source population of the cases without any matching, so that cases and controls labelled with the same ordinal number above are not related to each other,

   (b) by choosing for each case patient an individual control subject with the same age, and gender, such that each control is matched with the case having the same ordinal number above.

2. What appears to be the consequence to the rate ratio estimate here, if matching was applied in collecting the data but ignored in the analysis?

3.10 Cohort study and SMR

An occupational cohort study was started to estimate cancer mortality among male employees having a history of been working in a certain industry I during a certain time period, comparing it with that in a reference population which comprised economically active males at the same socioeconomic level living in the same area but not working in industry I. The results are displayed in the table on the next page. Calculate the following quantities:

1. Age-specific mortality rates in both populations and their ratios between the I-employees and the reference population. Does the rate ratio appear heterogenous over the age groups?

2. Crude mortality rates in the two populations and their ratio.

3. Age-adjused summary estimate of the rate ratio, using `glm` with Poisson error as indicated in the lectures.

4. Standardised mortality ratio (SMR).

5. Standardised mortality rates in the populations and their ratio using the reference population as the standard.

6. Are the rate ratio estimates sensitive to the choice of standard population?

7. Is there effect modification by age?

8. Is age a confounder in these analyses?

	Employees in I		Reference population	
Age group	Deaths	Person-years	Deaths	Person years
30–39	11	10,000	15	30,000
40–49	15	6,000	60	50,000
50–59	10	2,000	150	70,000
Total	36	18,000	225	150,000

3.11 Trial of tolbutamide

The effect of treating middle-aged and elderly diabetic subjects with a drug called tolbutamide vs. placebo as investigated in a famous randomised clinical trial (University Group Diabetes Program 1970). During a fixed follow-up period of 5 years with no losses, 30 out of the 204 patients randomised to tolbutamide died, and 21 out of the 215 patients in the placebo group died, too.

1. Calculate the following quantities:

(a) Incidence proportions (cumulative incidences) of death in both groups.

(b) Estimate of the risk ratio with its approximate 95% confidence interval between tolbutamide and placebo.

(c) Estimate of the risk difference and its approximate 95% confidence interval between tolbutamide and placebo.

2. Is tolbutamide dangerous to diabetics?

Chapter 4

Basic concepts in survival and demography

The following is a summary of relations between various quantities used in analysis of follow-up studies. They are ubiquitous in the analysis and reporting of results. Hence it is important to be familiar with all of them and the relation between them.

4.1 Probability

Survival function:

$$\begin{aligned} S(t) &= \mathrm{P}\{\text{survival at least till } t\} \\ &= \mathrm{P}\{T > t\} = 1 - \mathrm{P}\{T \leq t\} = 1 - F(t) \end{aligned}$$

where T is the variable "time of death"

Conditional survival function:

$$\begin{aligned} S(t|t_{\text{entry}}) &= \mathrm{P}\{\text{survival at least till } t| \text{ alive at } t_{\text{entry}}\} \\ &= S(t)/S(t_{\text{entry}}) \end{aligned}$$

Cumulative distribution function of death times (cumulative risk):

$$\begin{aligned} F(t) &= \mathrm{P}\{\text{death before } t\} \\ &= \mathrm{P}\{T \leq t\} = 1 - S(t) \end{aligned}$$

Density function of death times:

$$f(t) = \lim_{h \to 0} \mathrm{P}\{\text{death in } (t, t+h)\}/h = \lim_{h \to 0} \frac{F(t+h) - F(t)}{h} = F'(t)$$

Intensity:

$$\begin{aligned}\lambda(t) &= \lim_{h\to 0} \mathrm{P}\{\text{event in } (t, t+h] \mid \text{alive at } t\}/h \\ &= \lim_{h\to 0} \frac{F(t+h)-F(t)}{S(t)h} = \frac{f(t)}{S(t)} \\ &= \lim_{h\to 0} -\frac{S(t+h)-S(t)}{S(t)h} = -\frac{\mathrm{d}\log S(t)}{\mathrm{d}t}\end{aligned}$$

The intensity is also known as the hazard function, hazard rate, mortality/morbidity rate or simply "rate".

Note that f and λ are *scaled* quantities, they have dimension time^{-1}.

Relationships between terms:

$$\begin{aligned}-\frac{\mathrm{d}\log S(t)}{\mathrm{d}t} &= \lambda(t) \\ &\Updownarrow \\ S(t) &= \exp\left(-\int_0^t \lambda(u)\,\mathrm{d}u\right) = \exp\bigl(-\Lambda(t)\bigr)\end{aligned}$$

The quantity $\Lambda(t) = \int_0^t \lambda(s)\,\mathrm{d}s$ is called the *integrated intensity* or the **cumulative rate**. It is *not* an intensity (rate), it is dimensionless, despite its name.

$$\lambda(t) = -\frac{\mathrm{d}\log(S(t))}{\mathrm{d}t} = -\frac{S'(t)}{S(t)} = \frac{F'(t)}{1-F(t)} = \frac{f(t)}{S(t)}$$

The cumulative *risk* of an event (to time t) is:

$$F(t) = \mathrm{P}\{\text{Event before time } t\} = \int_0^t \lambda(u)S(u)\,\mathrm{d}u = 1 - S(t) = 1 - \mathrm{e}^{-\Lambda(t)}$$

For small $|x|$ (< 0.05), we have that $1 - \mathrm{e}^{-x} \approx x$, so for small values of the integrated intensity:

$$\text{Cumulative risk to time } t \approx \Lambda(t) = \text{Cumulative rate}$$

4.2 Statistics

Likelihood contribution from follow up of one person:

The likelihood from a number of small pieces of follow-up from one individual is a product of conditional probabilities:

$$\begin{aligned}\mathrm{P}\{\text{event at } t_4 | \text{entry at } t_0\} &= \mathrm{P}\{\text{survive } (t_0,t_1) | \text{ alive at } t_0\} \times \\ &\quad \mathrm{P}\{\text{survive } (t_1,t_2) | \text{ alive at } t_1\} \times \\ &\quad \mathrm{P}\{\text{survive } (t_2,t_3) | \text{ alive at } t_2\} \times \\ &\quad \mathrm{P}\{\text{event at } t_4 | \text{ alive at } t_3\}\end{aligned}$$

Each term in this expression corresponds to one *empirical rate*[1]
$(d, y) = (\#\text{deaths}, \#\text{risk time})$, i.e. the data obtained from the follow-up of one person in the interval of length y. Each person can contribute many empirical rates, most with $d = 0$; d can only be 1 for the *last* empirical rate for a person.

Log-likelihood for one empirical rate (d, y):

$$\ell(\lambda) = \log\big(\mathrm{P}\{d \text{ events in } y \text{ follow-up time}\}\big) = d\log(\lambda) - \lambda y$$

This is under the assumption that the rate (λ) is constant over the interval that the empirical rate refers to.

Log-likelihood for several persons. Adding log-likelihoods from a group of persons (only contributions with identical rates) gives:

$$D\log(\lambda) - \lambda Y,$$

where Y is the total follow-up time ($Y = \sum_i y_i$), and D is the total number of failures ($D = \sum_i d_i$), where the sums are over individuals' contributions with the *same* rate, λ, for example from the same age-class fro all individuals.

Note: The Poisson log-likelihood for an observation D with mean λY is:

$$D\log(\lambda Y) - \lambda Y = D\log(\lambda) + D\log(Y) - \lambda Y$$

The term $D\log(Y)$ does not involve the parameter λ, so the likelihood for an observed rate (D, Y) can be maximized by pretending that the no. of cases D is Poisson with mean λY. But this does *not* imply that D follows a Poisson-distribution. It is entirely a likelihood based computational convenience. Anything that is not likelihood based is not justified.

A linear model for the log-rate, $\log(\lambda) = X\beta$ implies that

$$\lambda Y = \exp\big(\log(\lambda) + \log(Y)\big) = \exp\big(X\beta + \log(Y)\big)$$

Therefore, in order to get a linear model for $\log(\lambda)$ we must require that $\log(Y)$ appear as a variable in the model for $D \sim (\lambda Y)$ with the regression coefficient fixed to 1, a so-called *offset*-term in the linear predictor.

4.3 Competing risks

Competing risks: If there are more than one, say 3, causes of death, occurring with (cause-specific) rates λ_1, λ_2, λ_3, that is:

$$\lambda_c(a) = \lim_{h\to 0} \mathrm{P}\{\text{death from cause } c \text{ in } (a, a+h] \mid \text{alive at } a\}/h, \quad c = 1, 2, 3$$

The survival function is then:

$$S(a) = \exp\left(-\int_0^a \lambda_1(u) + \lambda_2(u) + \lambda_3(u)\,\mathrm{d}u\right)$$

[1] This is a concept coined by BxC, and so is not necessarily generally recognized.

because you have to escape all 3 causes of death. The probability of dying from cause 1 before age a (the cause-specific cumulative risk) is:

$$F_1(a) = \mathrm{P}\{\text{dead from cause 1 at } a\} = \int_0^a \lambda_1(u)S(u)\,\mathrm{d}u \neq 1 - \exp\left(-\int_0^a \lambda_1(u)\,\mathrm{d}u\right)$$

The term $\exp(-\int_0^a \lambda_1(u)\,\mathrm{d}u)$ is sometimes referred to as the "cause-specific survival", but it does not have any probabilistic interpretation in the real world. It is the survival under the assumption that only cause 1 existed and that the mortality rate from this cause was the same as when the other causes were present too.

Together with the survival function, the cause-specific cumulative risks represent a classification of the population at any time in those alive and those dead from causes 1, 2 and 3 respectively:

$$1 = S(a) + \int_0^a \lambda_1(u)S(u)\,\mathrm{d}u + \int_0^a \lambda_2(u)S(u)\,\mathrm{d}u + \int_0^a \lambda_3(u)S(u)\,\mathrm{d}u, \quad \forall a$$

Subdistribution hazard Fine and Gray defined models for the so-called subdistribution hazard, $\tilde{\lambda}_i(a)$. Recall the relationship between between the hazard (λ) and the cumulative risk (F):

$$\lambda(a) = -\frac{\mathrm{d}\log\big(S(a)\big)}{\mathrm{d}a} = -\frac{\mathrm{d}\log\big(1 - F(a)\big)}{\mathrm{d}a}$$

When more competing causes of death are present the Fine and Gray idea is to use this transformation to the cause-specific cumulative risk for cause 1, say:

$$\tilde{\lambda}_1(a) = -\frac{\mathrm{d}\log\big(1 - F_1(a)\big)}{\mathrm{d}a}$$

Here, $\tilde{\lambda}_1$ is called the subdistribution hazard; as a function of $F_1(a)$ it depends on the survival function S, which depends on *all* the cause-specific hazards:

$$F_1(a) = \mathrm{P}\{\text{dead from cause 1 at } a\} = \int_0^a \lambda_1(u)S(u)\,\mathrm{d}u$$

The subdistribution hazard is merely a transformation of the cause-specific cumulative risk. Namely the same transformation which in the single-cause case transforms the cumulative risk to the hazard. It is a mathematical construct that is not interpretable as a hazard despite its name.

4.4 Demography

Expected residual lifetime: The expected lifetime (at birth) is simply the variable age (a) integrated with respect to the distribution of age at death:

$$\mathrm{EL} = \int_0^\infty a f(a)\,\mathrm{d}a$$

where f is the density of the distribution of lifetime (age at death).

The relation between the density f and the survival function S is $f(a) = -S'(a)$, so integration by parts gives:

$$\mathrm{EL} = \int_0^\infty a\big(-S'(a)\big)\,\mathrm{d}a = -\Big[aS(a)\Big]_0^\infty + \int_0^\infty S(a)\,\mathrm{d}a$$

The first of the resulting terms is 0 because $S(a)$ is 0 at the upper limit and a by definition is 0 at the lower limit.

Hence the expected lifetime can be computed as the integral of the survival function.

The expected *residual* lifetime at age a is calculated as the integral of the *conditional* survival function for a person aged a:

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

Lifetime lost due to a disease is the difference between the expected residual lifetime for a diseased person and a non-diseased (well) person at the same age. So all that is needed is a(n estimate of the) survival function in each of the two groups.

$$\mathrm{LL}(a) = \int_a^\infty S_{\mathrm{Well}}(u)/S_{\mathrm{Well}}(a) - S_{\mathrm{Diseased}}(u)/S_{\mathrm{Diseased}}(a)\,\mathrm{d}u$$

Note that the definition of the survival function for a non-diseased person requires a decision as to whether one will consider non-diseased persons immune to the disease in question or not. That is whether we will include the possibility of a well person getting ill and subsequently die. This does not show up in the formulae, but is a decision required in order to devise an estimate of S_{Well}.

Lifetime lost by cause of death is using the fact that the difference between the survival probabilities is the same as the difference between the death probabilities. If several causes of death (3, say) are considered then:

$$\begin{aligned} S(a) = 1 &- \mathrm{P}\{\text{dead from cause 1 at } a\} \\ &- \mathrm{P}\{\text{dead from cause 2 at } a\} \\ &- \mathrm{P}\{\text{dead from cause 3 at } a\} \end{aligned}$$

and hence:

$$\begin{aligned} S_{\mathrm{Well}}(a) - S_{\mathrm{Diseased}}(a) = &\ \mathrm{P}\{\text{dead from cause 1 at } a|\text{Diseased}\} \\ &+ \mathrm{P}\{\text{dead from cause 2 at } a|\text{Diseased}\} \\ &+ \mathrm{P}\{\text{dead from cause 3 at } a|\text{Diseased}\} \\ &- \mathrm{P}\{\text{dead from cause 1 at } a|\text{Well}\} \\ &- \mathrm{P}\{\text{dead from cause 2 at } a|\text{Well}\} \\ &- \mathrm{P}\{\text{dead from cause 3 at } a|\text{Well}\} \end{aligned}$$

So we can conveniently define the lifetime lost due to cause 2, say, by:

$$\begin{aligned} \mathrm{LL}_2(a) = \int_a^\infty &\ \mathrm{P}\{\text{dead from cause 2 at } u|\text{Diseased \& alive at } a\} \\ &- \mathrm{P}\{\text{dead from cause 2 at } u|\text{Well \& alive at } a\}\,\mathrm{d}u \end{aligned}$$

These quantities have the property that their sum is the total years of life lost due to the disease:

$$\mathrm{LL}(a) = \mathrm{LL}_1(a) + \mathrm{LL}_2(a) + \mathrm{LL}_3(a)$$

The terms in the integral are computed as (see the section on competing risks):

$$\mathrm{P}\{\text{dead from cause 2 at } x | \text{Diseased \& alive at } a\} = \int_a^x \lambda_{2,\mathrm{Dis}}(u) S_{\mathrm{Dis}}(u)/S_{\mathrm{Dis}}(a)\,\mathrm{d}u$$

$$\mathrm{P}\{\text{dead from cause 2 at } x | \text{Well \& alive at } a\} = \int_a^x \lambda_{2,\mathrm{Well}}(u) S_{\mathrm{Well}}(u)/S_{\mathrm{Well}}(a)\,\mathrm{d}u$$