Practicals - measuring disease frequency and association

Epidemiological methods in medical research

16 January 2024

The Bissau study

In rural Guinea-Bissau, 5274 children under 7 months of age were visited two times at home, with an interval of approximately 6 months. Information about vaccination (BCG, DTP, measles vaccine) was collected at baseline and at second visit. Death during follow-up was also registered. Other children move away during follow-up or survive until the second visit ('censored'). The following variables in the dataset are relevant for the exercise:

- `id` child id.
- `fuptime` follow-up time (in days). Maximum is 183 days.
- `fupstatus` status at follow-up: censored or dead.
- `bcg` vaccination status at baseline: yes or no.

The aim of this exercise is to compute different descriptive statistics and compare them between vaccine groups. The exercise is divided into 2 independent parts:

- A: analysis of a small subset of the data "by hand".
- B: analysis of the full data with dedicated functions from a statistical software.

In practice one would mostly use part B. However it can be challenging to master both software and statistics at once, and this is why we advice you to start with part A, i.e. focus on the understanding instead of the programming.

Note: questions 9, 10, and 12 involve statistical models (Poisson regression, logistic regression, Kaplan Meier estimator) that have not been introduced yet in this course. Do not hesitate to ask for help if you are not familiar with them.

Part A: by hand calculation

To start, we consider the data from 10 subjects extracted from the dataset:
(`fuptime` contains the follow-up time in days and `fupstatus` the status at follow-up)

```
     id fuptime fupstatus bcg             id fuptime fupstatus bcg
     20     183  censored  no              1      65      dead yes
     25     147      dead  no             29     183  censored yes
     31     183  censored  no             30     183  censored yes
     59     183  censored  no             32     183  censored yes
    526     177      dead  no             33     183  censored yes
```

1. Fill the following tables with:
    - left table: the number of children who were lost to follow-up (i.e censored) or died
    - right table: the number of children, number of children who died, and number of person-day

    among all children and per vaccination group. You can use a pocket calculator/computer/phone to obtain the number of person-day.

```
      status
bcg    censored dead                  bcg   n   death person-day
  no     ?        ?                     no    ?   ?     ?
  yes    ?        ?                     yes   ?   ?     ?
  all    ?        ?                     all   ?   ?     ?
```

2. Estimate among all children, those with BCG vaccination, those without BCG vaccination:
    - the *183-day risk of death*
    - the *odds of the 183-day risk of death*
    - the *daily and yearly incidence rate of death* [1]

```
                     bcg no  bcg yes  bcg all
risk                   ?        ?        ?
odds                   ?        ?        ?
rate (person.day)      ?        ?        ?
rate (person.year)     ?        ?        ?
```

3. What does the point estimate of each metric (risk, odd, rate) indicate about bcg vaccine efficacy?

4. What are the limitations of this analysis, i.e., what prevent you from concluding about vaccine efficacy?

[1] using that there are 365.25 days in a year

We could apply the same approach to the whole dataset

```
   id fuptime fupstatus bcg                    id fuptime fupstatus bcg
    1      65      dead yes         ···      5271     173  censored  no
    2     161  censored yes         ···      5272     143  censored yes
    3     166  censored  no         ···      5273     148  censored  no
    4     166  censored yes         ···      5274     182  censored  no
```

counting the number of times `fupstatus` is `dead` and summing the values in `fuptime`:

```
bcg        n  death person-day
  no    1973     97     325258
  yes   3301    125     554929
```

5. Is it a valid approach to estimate the 183-day risk? The incidence rate?

6. Here are, in chronological order (w.r.t. study time), the first lines for the children in the non-vaccinated group:

```
  id fuptime fupstatus bcg
2645       6  censored  no
1415       8      dead  no
1739       9  censored  no
3364       9      dead  no
3817      12  censored  no
 266      13      dead  no
 549      15      dead  no
```

Use the risk-rate relationship (slide 30 in the lecture) to retrieve the Kaplan-Meier estimates of the risk:

```
library(survival)
e.KM <- survfit(Surv(fuptime,fupstatus=="dead") ~ bcg, data = bissau)
head(setNames(1-e.KM$surv,e.KM$time),6)
```

```
           6            8            9           12           13           15
0.0000000000 0.0005070994 0.0010141988 0.0010141988 0.0015218135 0.0020294283
```

Part B: using dedicated functions of a statistical software

We will now use a statistical software (here the R software) to analyze the dataset. You can download the dataset from the course webpage or directly load it into R using:

```
## load data
bissau <- read.table(
  file = "https://bozenne.github.io/doc/Teaching/bissau.txt",
  header=TRUE
)
## only keep relevant column
bissau <- bissau[,c("id","fuptime","fupstatus","bcg")]
## convert categorical variable from numeric to factor
bissau$id        <- as.factor(bissau$id)
bissau$fupstatus <- as.factor(bissau$fupstatus)
bissau$bcg       <- as.factor(bissau$bcg)
## overview of the data
str(bissau)
```

```
’data.frame’:        5274 obs. of  4 variables:
 $ id       : Factor w/ 5274 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ fuptime  : int  65 161 166 166 161 161 166 166 166 166 ...
 $ fupstatus: Factor w/ 2 levels "censored","dead": 2 1 1 1 1 1 1 1 1 1 ...
 $ bcg      : Factor w/ 2 levels "no","yes": 2 2 1 2 2 2 2 2 2 2 ...
```

Incidence rate

7. Use the R function `xtabs` to create a 2 by 3 table with the number of children, number of deaths, and the number of person-years at risk by BCG vaccination status. You should retrieve the table just before question 5.

8. Estimate the incidence rate in person-days and person-years for each BCG vaccination group.
   Evaluate incidence differences and ratio. What do you notice?

9. You will see in Day 5 that incidence rates can be estimated using a Poisson regression:

```
e.rate <- glm(fupstatus=="dead" ~ bcg, data = bissau,
               family = poisson(link = "log"), offset = log(fuptime))
cbind(estimate = exp(coef(e.rateR)), exp(confint(e.rateR)))
```

```
Waiting for profiling to be done...
                estimate        2.5 %       97.5 %
(Intercept) 0.0002982248 0.0002427451 0.0003615709
bcgyes      0.7553162786 0.5799959555 0.9865547965
```

Compare the results with question 8?
What would you conclude regarding the vaccine efficacy (assuming no confounding)?
What is the impact of removing the intercept (use `~0+bcg`) from the model?

183-day risk of death:

10. Use the 2 by 3 table to evaluate the risk in each BCG group and the corresponding relative risk. Compare your results with the logistic regression. What is wrong with this approach?

```
e.logit <- glm(fupstatus=="dead" ∼ bcg, data = bissau,
               family = binomial(link = "logit"))
cbind(estimate = exp(coef(e.logit)), exp(confint(e.logit)))
```

```
Waiting for profiling to be done...
              estimate      2.5 %     97.5 %
(Intercept) 0.05170576 0.04188753 0.06302974
bcgyes      0.76118570 0.58093794 1.00017958
```

11. What is the following code achieving?

```
bissauS <- aggregate(cbind(dY=1,dN=fupstatus=="dead")∼fuptime,
                       data = bissau[bissau$bcg == "no",], FUN = "sum")
bissauS$dY.lag <- c(0,bissauS$dY[1:(length(bissauS$dY)-1)])
bissauS$Y <- sum(bissauS$dY) - cumsum(bissauS$dY.lag)
bissauS$r <- 1-cumprod(1-bissauS$dN/bissauS$Y)
head(bissauS,8)
```

```
  fuptime dY dN dY.lag    Y            r
1       6  1  0      0 1973 0.0000000000
2       8  1  1      1 1972 0.0005070994
3       9  2  1      1 1971 0.0010141988
4      12  1  0      2 1969 0.0010141988
5      13  1  1      1 1968 0.0015218135
6      15  1  1      1 1967 0.0020294283
7      16  2  2      1 1966 0.0030446577
8      18  1  1      2 1964 0.0035522725
```

12. A convenient way to perform the calculations of question 11 is to use the `survfit` function from the survival package introduced in question 6.
    Can you interpret the software output?
    Would you conclude about a lower risk in the vaccinated group?

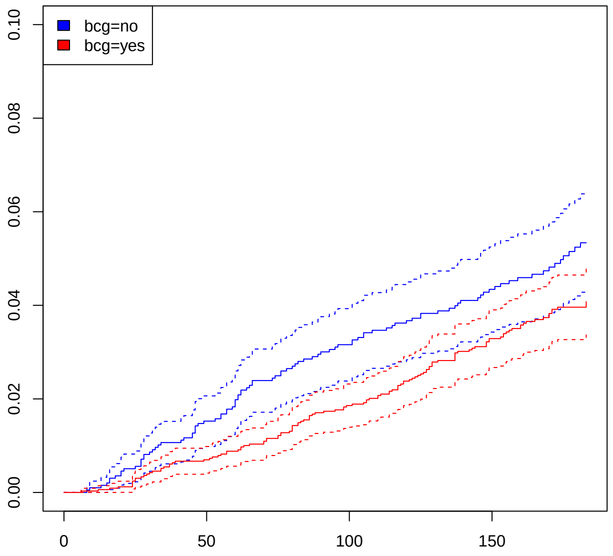
```
print(summary(e.KM, times = c(5,10,15,183)), digits = 4)
```

```
Call: survfit(formula = Surv(fuptime, fupstatus == "dead") ~ bcg, data = bissau)

                bcg=no
 time n.risk n.event survival   std.err lower 95% CI upper 95% CI
    5   1973       0   1.0000 0.0000000       1.0000       1.0000
   10   1969       2   0.9990 0.0007168       0.9976       1.0000
   15   1967       2   0.9980 0.0010137       0.9960       1.0000
  183    935      93   0.9466 0.0053654       0.9362       0.9572

                bcg=yes
 time n.risk n.event survival   std.err lower 95% CI upper 95% CI
    5   3299       0   1.0000 0.0000000       1.0000       1.0000
   10   3296       1   0.9997 0.0003032       0.9991       1.0000
   15   3295       1   0.9994 0.0004287       0.9986       1.0000
  183   1615     123   0.9592 0.0036162       0.9522       0.9663
```

```
plot(e.KM, fun = "event", ylim = c(0,0.1),
     conf.int = TRUE, col = c("blue","red"))
legend(x = "topleft", fill = c("blue","red"),
       legend = sort(unique(bissau$bcg)))
```


🟦 bcg=no
🟥 bcg=yes