Practicals - measuring disease frequency and association

Epidemiological methods in medical research

19 January 2023

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. R users will find in section 6.5 of the document `L2-summary.pdf` useful R syntax to answer these questions.

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
20	183	censored	no
25	147	dead	no
31	183	censored	no
59	183	censored	no
526	177	dead	no

id	fuptime	fupstatus	bcg
1	65	dead	yes
29	183	censored	yes
30	183	censored	yes
32	183	censored	yes
33	183	censored	yes

1. Fill the following tables with the number of children of children who were lost to follow-up (i.e censored) or died by vaccination group (left table) and the number of children, number of children who died, and number of person-day by vaccination group. You can use a pocket calculator/computer/phone to obtain the number of person-day.

	status	
bcg	censored	dead
no	?	?
yes	?	?
total	?	?

bcg	n	death	person-day
no	?	?	?
yes	?	?	?
total	?	?	?

2. Estimate for children with or without BCG vaccinations:

   - the *183-day risk of death*
   - the *odd of the 183-day risk of death*
   - the *daily and yearly incidence rate of death* [1]

	bcg no	bcg yes	bcg total
risk	?	?	?
odd	?	?	?
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 limitation 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
       1      65      dead yes
       2     161  censored yes
       3     166  censored  no
       4     166  censored yes
       5     161  censored yes
---
    5270     183  censored  no
    5271     173  censored  no
    5272     143  censored yes
    5273     148  censored  no
    5274     182  censored  no
```

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

```
t23 <- xtabs(cbind("n" = 1,
               "death" = fupstatus=="dead",
          "person-day" = fuptime) ~ bcg,
             data = bissau)
t23
```

```
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 vaccinated group. Can you evaluate the 5-, 10-, and 15-day risk of death in that group?

```
       id fuptime fupstatus bcg
2876 2876       2  censored yes
89     89       4  censored yes
1908 1908       5  censored yes
2551 2551       6      dead yes
2786 2786       9  censored yes
1344 1344      12      dead yes
598   598      15  censored yes
3736 3736      16      dead yes
```

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. Make a 2 by 3 table with the number children, number of deaths, and the number of person-years at risk by BCG vaccination status (i.e. retrieve the table just before question 5). Estimate the incidence rate per BCG vaccination group.

8. Evaluate incidence differences and ratio with their 95% confidence limits. R users can use the `effx` function from the Epi package.
   What would you conclude?

The incidence rate can also be obtained using a Poisson regression model (`proc genmod` in `SAS` and `glm` in `R`), using log(person-years) as 'offset', a 'log link' function, and exponentiate the resulting estimates. Alternatively the Epi package has a family `"poisreg"` that allows a more natural specification of the Poisson model, see `?poisreg` after loading the Epi package with `library(Epi)`.

9. Compute the incidence rate (per person.day and person.year) in the two BCG vaccination groups with its confidence interval using a Poisson regression model. What is the impact of including or not an intercept in 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. To get confidence intervals R users can use the `effx` function from the Epi package.
    Equivalently one can use a logistic regression (`proc genmod` in `SAS` and `glm` in `R`), using `fupstatus` as an outcome and `group` as covariate.
    What is wrong with this approach?

An appropriate analysis would be based on a Kaplan-Meier estimator

11. The dataset shown at the end of question 6 can be obtained with the commands below. Can you estimate the 5-, 10-, 15-day risk using basic operations with your statistical software? (R users may find the function `cumprod` convenient)

```
bissau.order <- bissau[order(bissau$fuptime),]
bissau.order8 <- bissau.order[bissau.order$bcg=="yes",][1:8,]
```

12. Use a dedicated function in your statistical software to obtain the Kaplan Meier estimator of the risk over time. R users can use the function `survfit` from the survival package, SAS users can use the `proc lifetest`. Extract the estimated risk at 5-, 10-, 15-, and 183- days and compare it to previous results.
    Can you also get a confidence interval of for the risk and risk difference?