Basic knowledge of measurement and epidemiological research

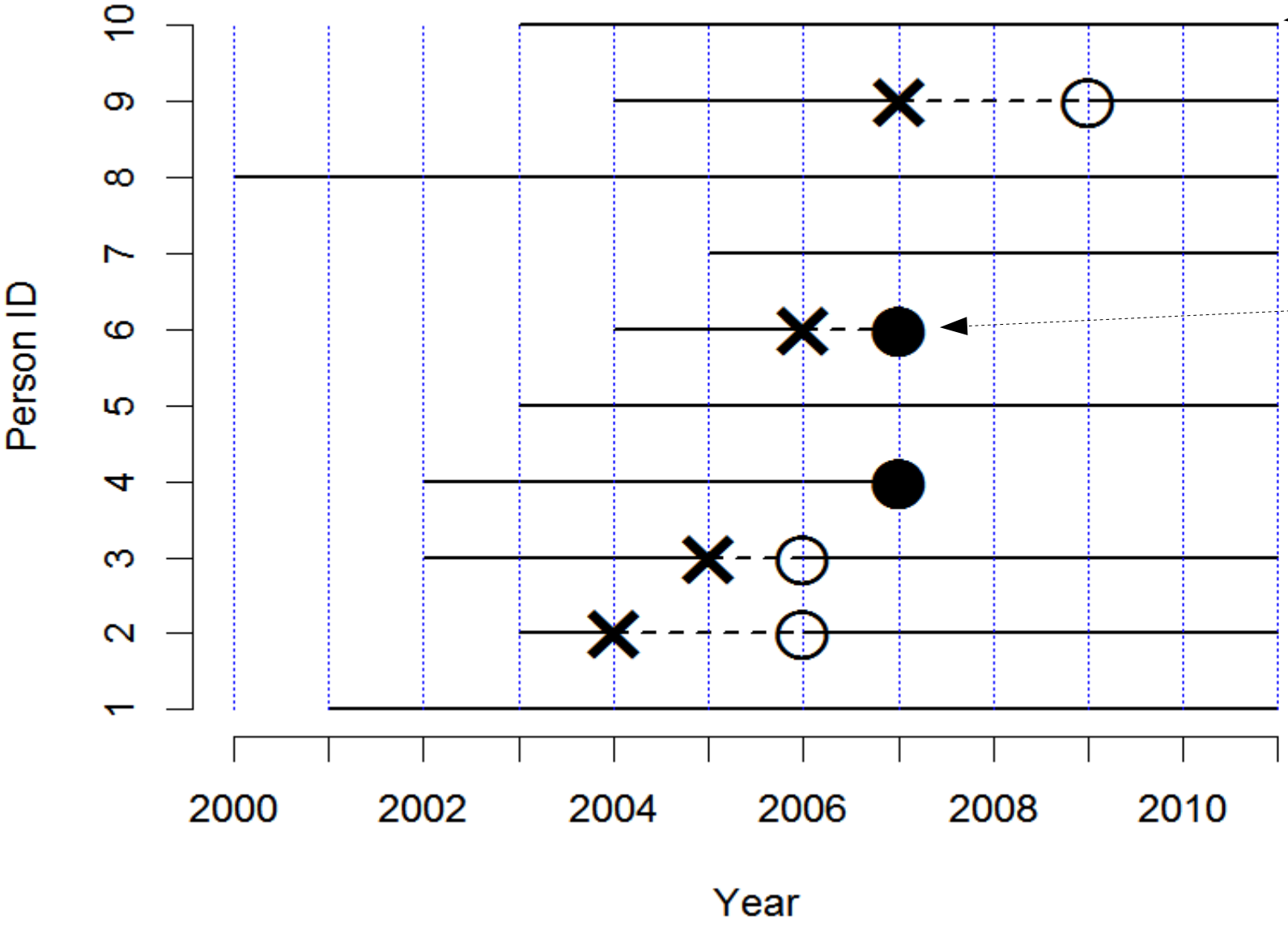
- Validity
  - To correctly scale the measurement of which the researcher truly wants to measure.
  - eg. ELISA test using the antibody with low specificity against the target molecule has low validity, because other molecules are included in the measurement.
- Accuracy
  - Small bias (systematic error)
  - eg. Measurement without zero-adjustment in advance has low accuracy.
- Precision
  - Low stochastic (random) error
  - Same as narrow confidence intervals and small CV.
  - Measurement with low sensitivity has low precision.

As the result of typical epidemiologic study ...

- Cross tabulation
  - Evaluating the degree of association between the 2 categorical variables.
- In epidemiologic study, the association between [disease/health] and [exposed/nonexposed].
  - How to measure the amount of disease
    - Cross-sectional or Case-control study→prevalence or odds
    - Cohort study→risk or incidence rate
  - How to evaluate the degree of association (=effect) between disease status and exposure.
    - Difference or ratio
  - In hypothesis testing, what is the meaning of "p-value is less than the significance level (eg. 0.05)".
    - Non-opportunistic association or difference = statistically significant
    - "Not significant" result does not support no-association nor no-difference.
    - (Cons.) Original information is shrinked to binary. (Pros.) Useful for judgement.
    - Recently showing confidence intervals is preferred than p-value.

Measuring the amount of disease

- Let's see the hypothetical situation of disease occurrence. The graph below shows: each horizontal line = observation of each individual aligned with year, solid line = healthy, dashed line = in disease, × = incidence, ○ = recovery, ● = death.

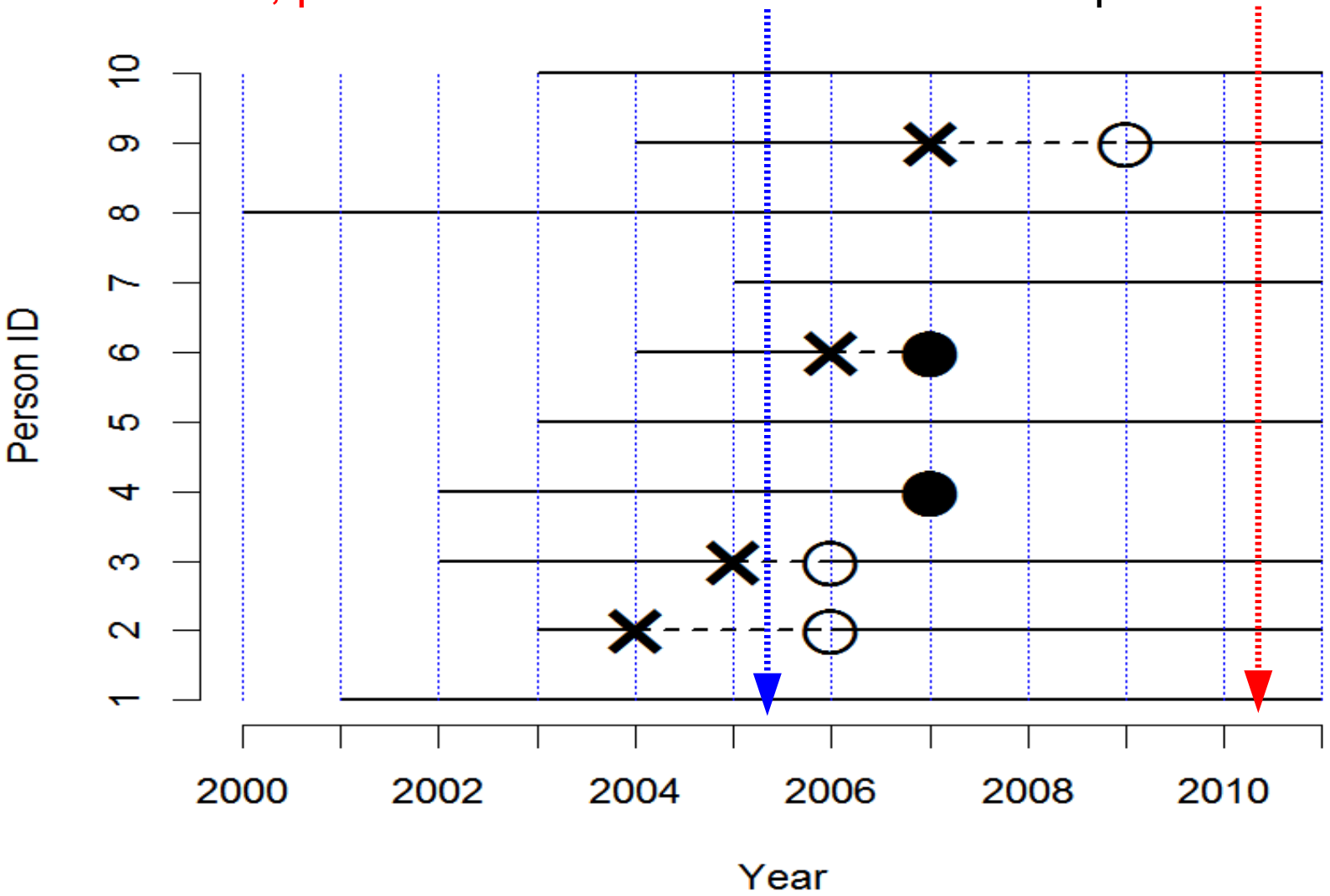


Born in the beginning of 2003 and lived healthy until the end of 2010.

Born in the beginning of 2004, suffering from the target disease in the beginning of 2006, then died in the early 2007.

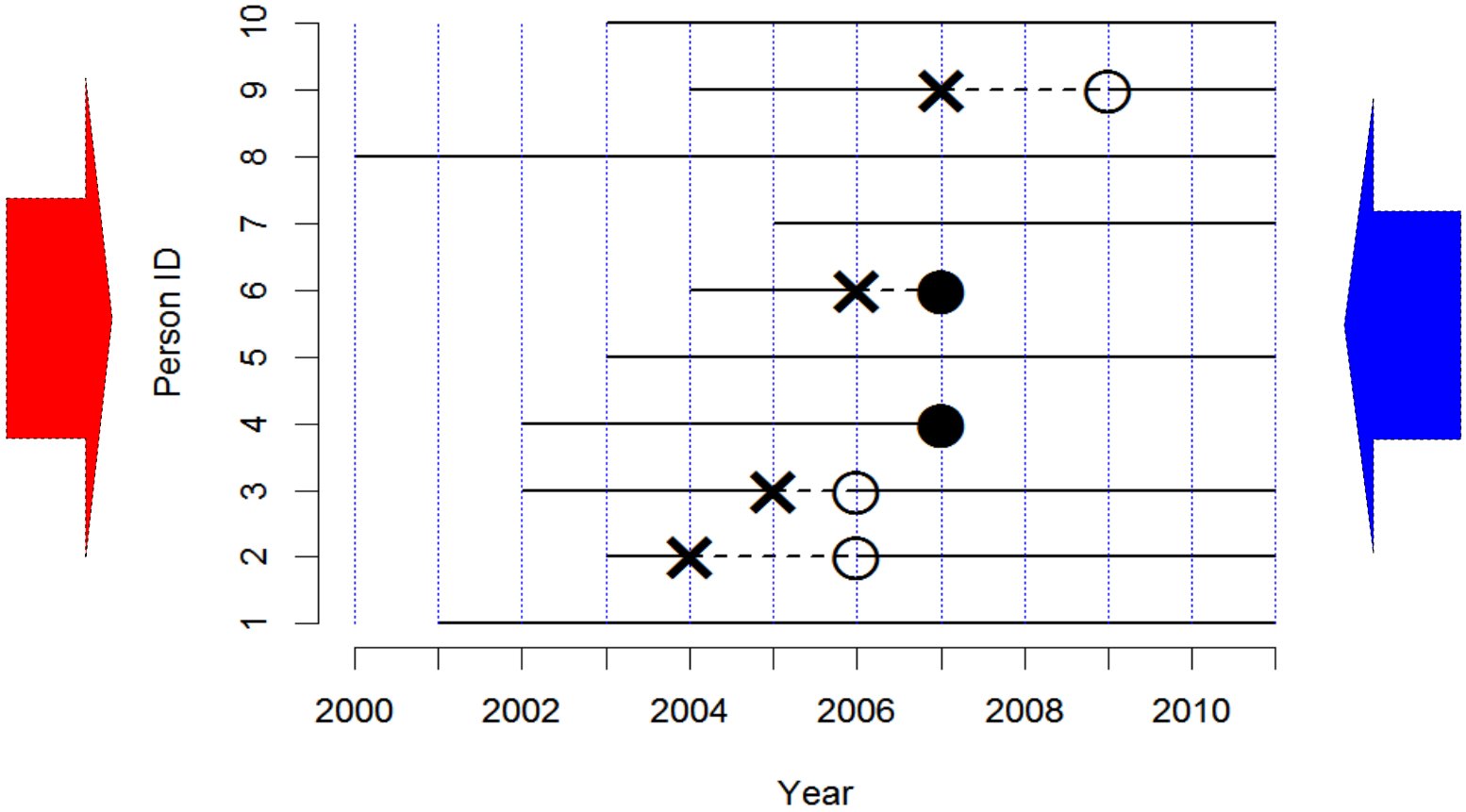
Prevalence

- If the researcher conduct cross-sectional study in mid-2005, 2 disease patients among 10 people are counted, so that the prevalence is 2/10 = 0.2 (Disease odds (the ratio of disease patients to healthy people) is 2/8=0.25)←Research is easy.
- However, prevalence is 0 in mid-2010←Low representativeness



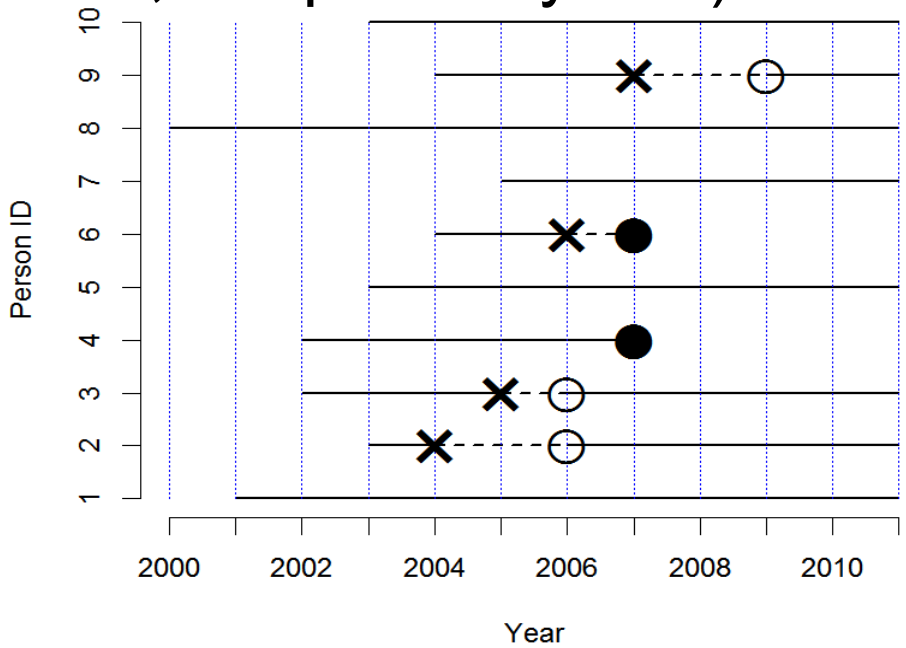
Risk (cumulative incidence)

- For the surviving 8 children in the beginning of 2011, ask their guardians to recall the past incidence of the target disease. Three of them had the experience of suffering from that disease, then we can calculate the risk during 11 years until 2011, as 3/8 = 0.375←(Bad method) Easy and cheap, but already died children were missed. If the complete record of birth, incidence and death exist, retrospective cohort study is possible, but it's rare.
- From the beginning of 2000, if the researcher follow the 10 children for 11 years to record birth, disease incidence, recovery and death (cohort study), 4 of 10 children suffered from the target disease in 11 years. Thus the risk during 11 years until 2011 is 4/10 = 0.4. However, the risk for a year after the birth is 1/10=0.1←Risk depends on observation period.



Incidence rate

- The cohort study from 2000 provides whole person-years data.
- If everybody can suffer from the target disease at most once during the lifetime, the patients lose susceptibility to that disease and thus the one is removed from population at risk. Incidence rate is, the number of disease occurrence divided by the sum of the susceptible person-years. The dimension is 1/year.
- If everybody can suffer from the target disease more than once, the incidence rate of the population is the number of incidence of that year divided by the population at risk on the mid-day of that year (usually per 100,000 person-years).

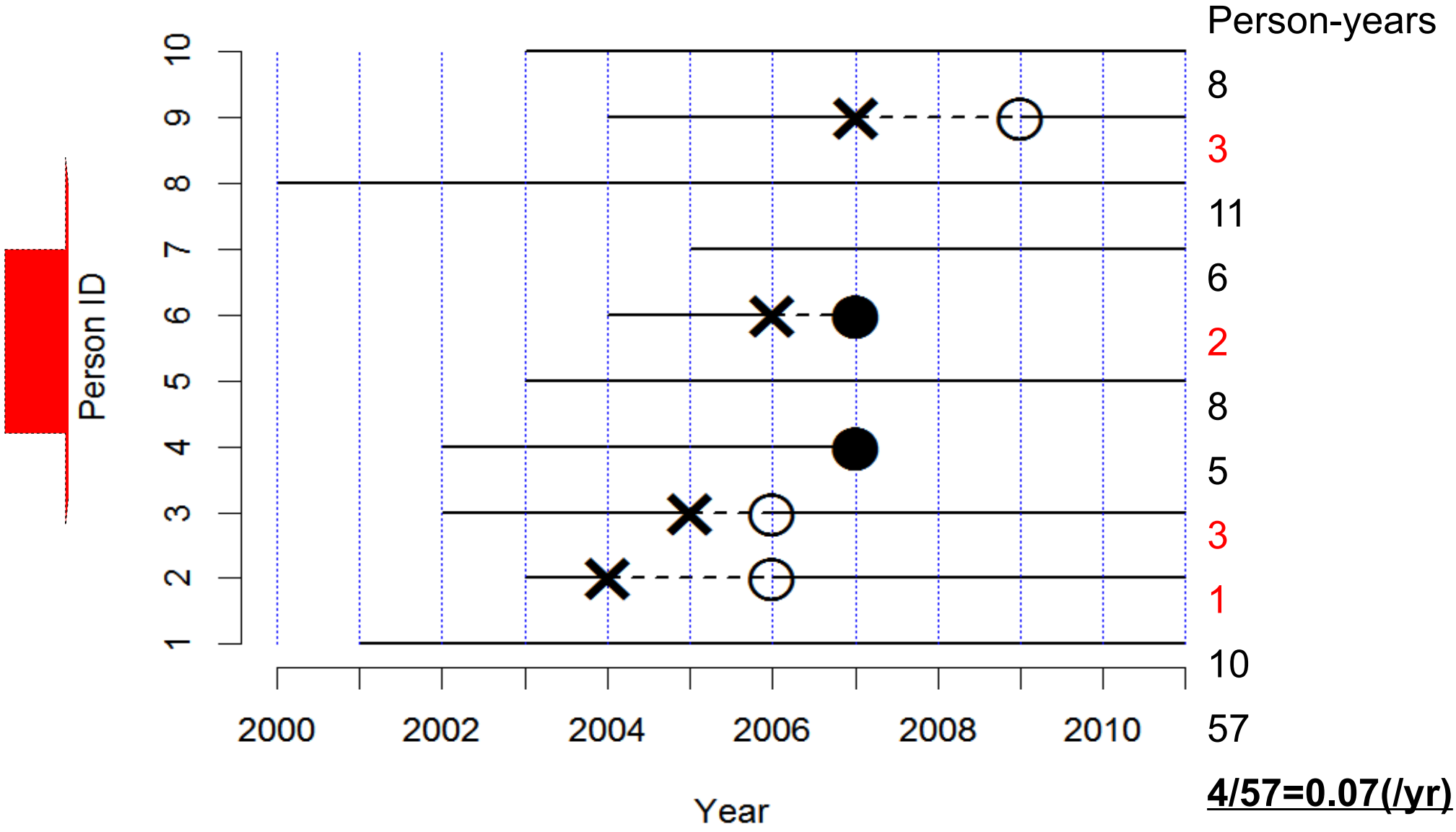


How's the incidence rate in the left graph?

↓

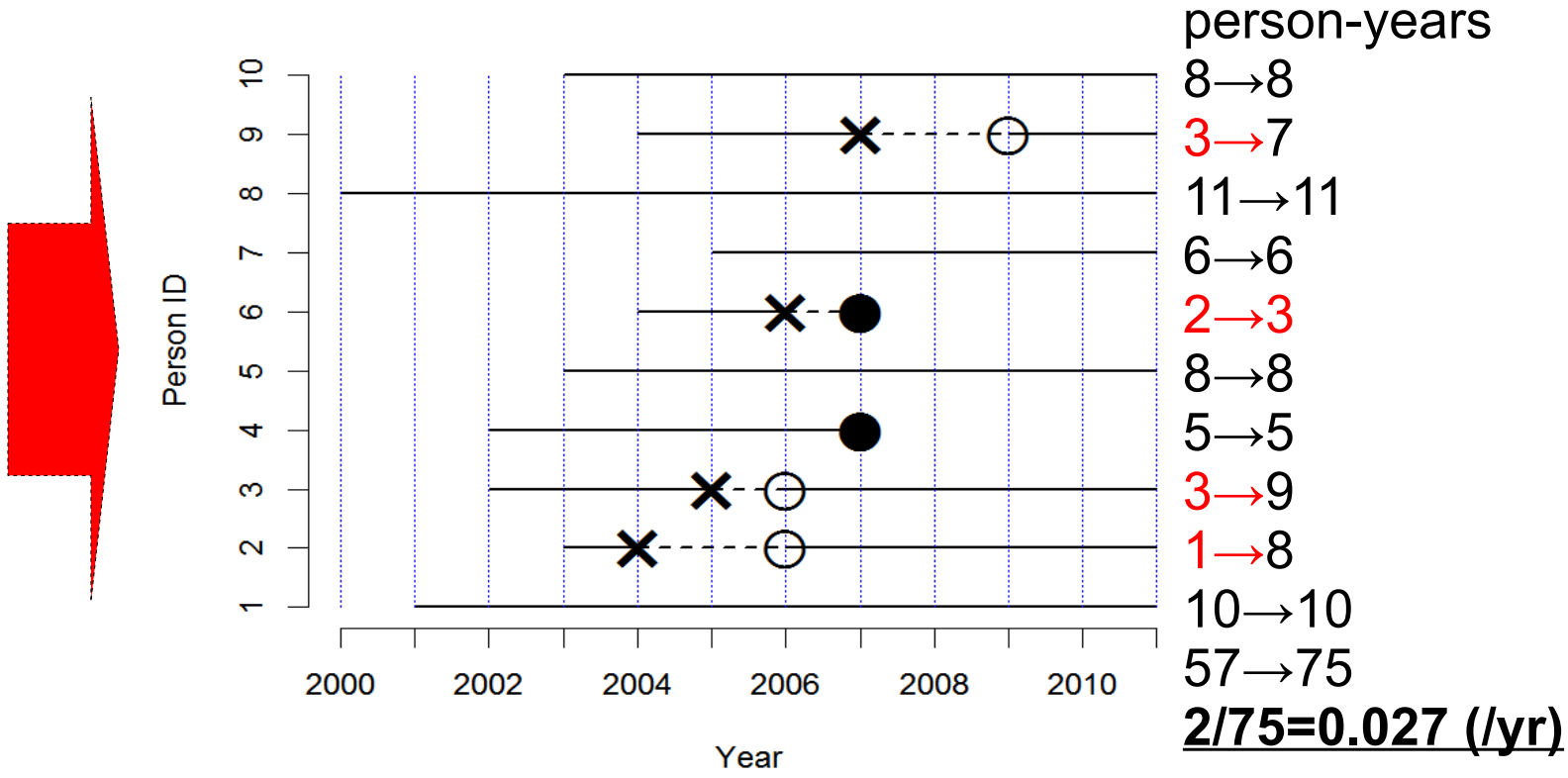
Answer is shown in the next slide

Example of incidence rate calculation based on cohort study



Mortality rate

- If we use death instead of incidence of the disease as the endpoint of observation, we can get mortality rate instead of incidence rate → 0.027/yr as shown below
- Death is considered as same as the disease which can occur at most once during the lifetime.
- For the large-size population, annual number of death divided by the mid-year day's population→In the example below, 0.2/yr in 2007, 0/yr in other years.



Association between exposure and disease = Comparison of disease amount between exposed and non-exposed groups

- Typical comparisons
  - Difference (absolute) … Showing public health impact
  - Ratio (relative) … Assessing the strength of causal relation
  - Each has specific mean
- By the measurement of disease amount
  - Risk→Difference (RD) or Ratio (RR)
  - Incidence rate→Difference (IRD) or Ratio (IRR)
  - Mortality rate→Difference (MRD) or Ratio(MRR)
  - Prevalence→Odds Ratio

Absolute comparison: RD, IRD ＝Attributable risk (= Excess risk)

- (Hypothetical example) Follow up 100,000 residents living close to high voltage cable for 5 years, then 2 leukemia patients are found every year. Similarly follow up 100,000 residents living apart from high voltage cable, then 1 leukemia patient is found every year. No difference except proximity to high voltage cable between 2 groups.
- RD = 10/100000-5/100000=5/100000 (=5e-5)
- IRD = 10/(100000+99998+99996+99994+99992) -5/(100000+99999+99998+99997+99996) ≒0.0000100006（/year）
- Difference looks small due to small risk/incidence rate.
- Using R and fmsb package, when you type library(fmsb); riskdifference(10, 5, 100000, 100000); ratedifference(10, 5, 499980, 499990) Then you get the point estimates and 95% confidence intervals of RD and IRD.

Relative comparison (1): RR, IRR

- Same example
- RR = (10/100000)/(5/100000)=2
- IRR = $\frac{10/(100000+99998+99996+99994+99992)}{5/(100000+99999+99998+99997+99996)} \fallingdotseq 2$
- Both means "Living close to high voltage cable raises the leukemia risk twice"
- Statistical significance
  - Testing the null-hypothesis "ratio is 1"
  - Calculation is easy if we use the software like R (EZR), SAS, JMP.→ Since p-value is slightly less than 0.2, the result is not statistically significant.
  - By R with fmsb package, library(fmsb); riskratio(10, 5, 100000, 100000); rateratio(10, 5, 499980, 499990) Then you get the point estimates with 95% confidence intervals.

Relative comparison (2): OR

- Disease amount in cross-sectional study is measured by prevalence. However, difference or ratio of prevalence doesn't make sense. Instead, odds ratio (OR) is calculated as the disease odds of people with specific attributes divided by the disease odds of people without that attributes.
- OR in case-control study is exposure odds in patients divided by exposure odds in controls.
  - Cohort study of 200,000 people for 5 years may detect only 15 cases in rare disease. Such study has very low statistical power and efficiency is low.
  - Case-control study of 100 patients of rare disease (eg. child leukemia) in a specific hospital and 200 controls (eg. injury) in the same hospital with information of exposure is highly effective. Both 20 people exposed (eg. living close to high voltage cable)

	Cases	Controls
Exposed	20	20
Nonexposed	80	180

* OR=(20/80)/(20/180)=2.25

fisher.test(matrix(c(20,80,20,180),2)) in R resulted in $p<0.05$

Of course, by using fmsb package, library(fmsb); oddsratio(20, 80, 20, 180) will give you the point estimates and 95% confidence intervals of OR.