Epidemiology (4)
Chapter 4. Measuring disease occurrence

Minato NAKAZAWA, Ph.D.
<minato-nakazawa@people.kobe-u.ac.jp>

Measures of Disease Occurrence, Association and Causal Effects

- Measurement is a central feature of epidemiology
  - As the study of the occurrence of illness
- Broad interpretation of illness
  - Injuries
  - Birth defects
  - Health outcomes
  - Other health related events and conditions
- As the measures of disease occurrence, 3 fundamental ones are introduced in this chapter
  - Risk (~ Incidence Proportion)
  - Incidence rate
  - Prevalence (~ Prevalence Proportion)

Risk (~ Incidence Proportion)

- Risk (~ Incidence proportion)
  - Probability
  - In a small group, **the probability** that **a person** will develop a given disease
  - In larger groups, let A the number of people who develop the disease among the number of people in the beginning of the observation N, **risk** can be defined as **A/N** (A < N, thus A/N < 1)
    - It assumes that all of the N people are followed for the entire period
    - The **average risk for a group** is also referred as **incidence proportion**
- Advantage of using risk
  - Readily understood by many people
- To note, What does "60 years old women have 2% risk of dying of cardiovascular disease" in newspaper mean?
  - Within next 24 hours, it's not true
  - Even for 1 year, it's too high
  - The length of time over which the risk applies is essentially important. Without the observation period, risk has no meaning

Table 4.1. Comparison of incidence proportion (risk) and incidence rate

Property	**Incidence Proportion**	**Incidence Rate**
Smallest value	0	0
Greatest value	1	Infinity
Units (dimensionality)	None	1/time
Interpretation	Probability	Inverse of waiting time

- The risk increases with time (See, Figure 4.1)
  - Pattern A: Risk climbs rapidly early to reach plateau
  - Pattern B: Risk climbs at slowly but steadily rate
- Possible conditions
  - A: Susceptible people in the beginning become immunized after recovery
  - B: Exposure to DES in the beginning gradually develop vaginal adenocarcinoma / Chronic NCDs with aging

Pros and Cons of risk

- Cumulative measure
  - For a given person, risk increases with the length of observation
  - For a given period, risk**s** for a person can rise or fall with time
    - 1-year risk of dying in an automobile crash for a driver: For any one person, risk accumulates, but 1-year risk is greater for most drivers in their teenage years than 50s
- Drawback as a tool for assessing the disease occurrence
  - Impossible to measure risk over any appreciable time interval
  - During sufficient time, some people in the population will die from causes other than the outcome under study (**competing risks**)
    - (eg.) DV in 10000 married women over 30 years: Some of them will die before the completion of 30 years by cardiovascular disease, cancer, infection, vehicular injury, and other causes. If a woman dies by cancer after 5 years, we don't know whether the woman will become a victim of DV during the subsequent 25 years if she wouldn't suffer from cancer
    - If the number of person died due to competing risks is included in denominator N, the risk of DV will be underestimated
  - When we assess the all-causes mortality, no competing risks. When we assess the risk within short time period, the influence of competing risks is small
    - Evaluation of the efficacy of Salk vaccine for polio in schoolchildren in 1954 was done by 1-year follow up and during that period competing risks are negligible, it was reasonable
  - **Loss to follow-up** is another issue.

(Column) Attack rate and case fatality *rate* (Note: case fatality **risk** is preferred now)

- Attack rate: term for risk used in connection with infectious outbreaks
  - Risk of contracting a condition during an epidemic period
  - (eg.) When flu epidemic has 10% attack rate, 10% of the population will develop the disease during the epidemic
  - Time reference is not stated but implied by biological nature of the disease
    - Usually short, typically a couple of months, sometimes less
- Secondary attack rate: Attack rate among susceptible people who come into direct contact with primary cases (the cases infected in the initial wave of an epidemic) (See, Chapter 13)
- Case-fatality rate (it's older term, so hereafter I use **Case-Fatality Risk**): The proportion of people dying of the disease (fatality) among those who develop the disease (case). In general, the denominator is the number of **confirmed** cases. Sometimes, the numerator is the number of all deaths in the cases, but usually the number of deaths caused (succumbed) by that disease (See, Chapter 13).
  (cf.) Kelly H, Cowling BJ (2013) Case Fatality: Rate, Ratio, or Risk? *Epidemiology* 24(4): 622-623. https://doi.org/10.1097/EDE.0b013e318296c2b6
  - It indicates the severity of the disease.
  - Same as attack rate, time reference is usually implicit
  - CFR of measles in USA is 1.5/1000. CFR of COVID-19 in USA in March 2022 was 1.2/100. This time reference is much shorter than other outcomes than death by measles infection
  - For the diseases with long term process like MS, the interpretation of CFR is difficult and thus other measures (such as 5-year survival rate) may be used.

Incidence Rate

- The issues of competing risks can be addressed by changing the **denominator** from the population observed to the **total person-time observed**.
- Incidence rate = A/Time = (Number of subjects developing disease)/(Total time experienced for the subjects followed)
- The denominator is the sum of the time that each person is followed for every member. If 5 people are followed for 30 years, the denominator is 150 person-years. If among those 5 people, 4 people were followed for 30 years but 1 died after 5 years from the beginning of the observation, the denominator is 125 person-years.
- For people who don't die during follow-up, 2 methods of counting
  - If the disease can recur (like upper respiratory tract infection), the numerator includes all recurrence, the denominator includes all the time during which each person was at risk of getting infected
  - If the disease can occur only once (or outcome is death, in which the incidence rate is the mortality rate), the person ceases from the population at risk after the event occurrence
  - If a person is lost from follow-up, or dies from a competing risk, the person also ceases from the population at risk

Examples

- Hypothetical example of leukemia (Figure 4.2)
  - No value for time is given in the text, but here I assumed as follows
  - Among 5 people, only the 1st died of leukemia after 2 years. 2nd died of competing cause after 3 years. 3rd lost to followup after 1 year. 4th and 5th are censored (completed 5 years followup period without event)
- Incidence rate of leukemia death is 1/(2+3+1+5+5) = 1/16 (/year)
- Comparison between risk and incidence rate (Table 4.1)
  - Risk is probability, no unit, ranges [0, 1]
  - Incidence rate has dimensionality of 1/time, ranges from 0 to infinity

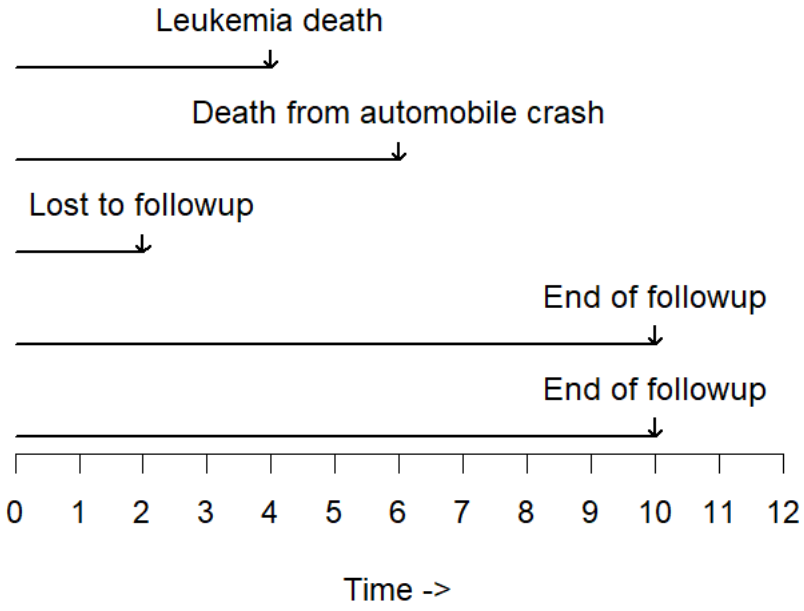

Figure 4.2. Time at risk for leukemia death for 5 people.

- Calculate an incidence rate in a population as 47 cases occurring in 158 person-months (=13.17 person-years).
  - 47 cases/158 person-months
  - 0.30 cases/person-months
  - 47 cases/13.17 person-years
  - 3.57 cases/person-years
- The change of unit results in the change of values

Annual incidence and waiting time

- Incidence rates commonly are described as **annual incidence**, in the form of "50 cases per 100,000"
  - Actually 50 cases per 100,000 person-years or 50/100000 yr^{-1}
  - Negative 1 in the exponent means inverse
- Different from risk, incidence rate may not easily understood
- Dimensionality of an incidence rate is that of the reciprocal of time
  - Under steady-state conditions, the reciprocal of incidence rate is **waiting time**, the average time until an event occurs
  - (eg.) If incidence rate is 3.57 cases per person-year, the waiting time is 1/3.57 = 0.28 years
  - (another eg.) A mortality rate of 11 deaths per 1000 person-years means the average waiting time until death is 1000/11=90.9 years, referred as expectation of life or expected survival time
    - Mortality rate changes with time, so that we cannot assume steady-state conditions. Taking reciprocal of mortality rate is not useful as the method to estimate expectation of life (Usually life table or survival analysis is used)

The Relation between Risk and Incidence Rate

- Converting the incidence rate measures to risk measures is convenient
  - The simplest formula: Risk ≈ Incidence rate x Time [4-1]
  - Confirming dimensionality is a good habit
    - Risk is a proportion, no dimension
    - Incidence rate [/time] x Time [time] → no dimension
  - Checking the range of measures is also useful
    - Risk ranges [0, 1]
    - Incidence rate ranges [0, infinity]
    - Time ranges [0, infinity]
    - Products of [0, infinity] and [0, infinity] ranges [0, infinity], does not always match with [0, 1]
  - Only for the risk<20%, equation [4-1] can work as approximation
- (eg.) When incidence rate of lung cancer is 8/10000 person-years and we followed the population for 1 year, risk is 8/10000 for 1 year period. If the same rate is applied for 0.5 year, the risk is 4/10000 for a half year period.
- (eg.2) If mortality rate is 11/1000 person-years and we followed the population for 20 years, the risk of death over 20 years will be 0.22 (22%). It means 220 deaths among 1000 people occurring within 20 years. However, it neglects the fact that the size of population at risk shrinks as deaths occur. Tables 4.2 (a kind of life table) shows 197 deaths instead of 220 deaths will occur within 20 years (It's refereed as **exponential decay**, as shown in Figure 4.3).

Simplified life table

- In a **hypothetical cohort** of 100,000 people **followed for 85 years**
  (Note: Practically, instead of long-term cohort, from annual age-specific mortality rate (mx), qx is calculated by n*mx/(1+n*(1-ax)*mx, where n is duration of each age-class and ax is fraction of last-year lived. From qx, ax and initial lx, the following table will be constructed.)
- Initial number at risk = 100,000
- Assume **no competing risk** (the number at risk at the start of each age group is reduced only by deaths from motor vehicle injury), **no lost to followup**.
- The qx is obtained by dx/lx for each age group
- The px is 1 – qx for each age group
- The cumulative survival probability is the product of px up to that age
- 1 minus final cumulative survival probability gives the total risk for 85 years
  - 1 – 0.98378 = 0.01622 (1.6%)

Table 4.3. Life table for death from motor vehicle injury from birth through age 85

Age (x)	Number at risk (lx)	Deaths in interval (dx)	Risk of dying (qx)	Survival probability (px)	Cumulative survival probability
0-14	100,000	70	0.00070	0.99930	0.99930
15-24	99,930	358	0.00358	0.99642	0.99572
25-44	99,572	400	0.00402	0.99598	0.99172
45-64	99,172	365	0.00368	0.99632	0.98807
65-84	98,807	429	0.00434	0.99566	0.98378

More realistic life table

(But still based on cohort study, in demography, usually the calculation starts from age-specific mortality. See, https://minato.sip21c.org/demography-special/demography-2020-06.pdf)

- In a hypothetical cohort of 100,000 people followed for 85 years
- Initial number at risk = 100,000
- Considering **competing risk** and **lost to followup** (censored).
- Effective number at risk = lx – (lost to followup / died of other causes)/2
  - Censoring is assumed to occur uniformly throughout each age interval.
- The qx = dx/lx' is approximately same as Table 4-3.
- The px is 1 – qx for each age group
- The cumulative survival probability is the product of px up to that age
- 1 minus final cumulative survival probability gives the total risk for 85 years
  - 1 – 0.98378 = 0.01622 (1.6%)

Table 4.4. Life table for death from motor vehicle injury from birth through age 85

Age (x)	At risk (lx)	MVI deaths in interval (dx)	Lost to followup / died of other causes	Effective number at risk (lx')	Risk of dying (qx)	Survival probability (px)	Cumulative survival probability
0-14	100,000	67	9,500	95,250	0.00070	0.99930	0.99930
15-24	90,433	301	12,500	84,183	0.00358	0.99642	0.99572
25-44	77,632	272	20,000	67,632	0.00402	0.99598	0.99172
45-64	57,630	156	30,000	42,360	0.00368	0.99632	0.98807
65-84	27,204	64	25,000	14,704	0.00435	0.99565	0.98377

Prevalence proportion

- Incidence proportion (risk) and incidence rate are measures that assess the frequency of disease onset.
- **Prevalence proportion** (often referred as prevalence) does not measure disease onset, but disease status.
  - Disease status: Considering disease as being either present or absent
  - Prevalence: Proportion (P/N) of people who have disease at a given time (P) in a population (N).
  - (eg.) On July 1, 2001, among 10,000 women residents of a town, 1,200 had hypertension. Prevalence of hypertension was 1200/10000 = 0.12 (12%).
- The factors affecting prevalence
  - Disease occurrence
  - Disease duration: Length of time that a person has disease
- Prevalence measures the **disease burden** on a population
- In a steady state, P/(1 – P) = ID. P/(1 – P) is known as the **prevalence odds**.
  - P: Prevalence
  - I: Incidence rate
  - D: Average duration of disease
- If prevalence is 0.75, the prevalence odds is 0.75/(1 – 0.75) = 3. If prevalence is 0.20, the prevalence odds is 0.20/(1 – 0.20)=0.25
  - For small prevalences (eg. <0.1), the value of the prevalence and the prevalence odds become close, P ≈ ID
- P/(1 – P) = ID <=> P = ID/(1+ID)
- Usually measured for public health administration, but it also can be measured for causal inference.

  (eg.) The proportion of infants who are born alive with a defect of ventricular septum of the heart (see, https://www.mayoclinic.org/diseases-conditions/ventricular-septal-defect/symptoms-causes/syc-20353495) is prevalence. Measuring the incidence rate or risk of ventricular septum defects needs ascertainment of a population of embryos at risk and measurement of the defect's occurrence: Usually impossible to get such data

(Column) Prevalence of characteristics

- Prevalence measures status
- Sometimes used to describe the status of characteristics or conditions other than disease in a population
- (eg.) The proportion of a population that engages in cigarette smoking often is described as the prevalence of smoking
- The proportion of a population exposed to a given agent is often referred to as the **exposure prevalence**
  - The number of exposed to a given agent (E) among total population (N) gives the exposure prevalence (E/N)
  - Similarly, **exposure odds** can be defined as E/(N – E)
- Prevalence can be used to describe the proportion of people in a population who have brown eyes, type O blood, or an active driver's license. For causation, it is sometimes useful.

MEASURES OF ASSOCIATION

- Three major objectives of epidemiologic studies: In all of these, quantitative measure of the extent of disease is needed. It's ***measures of association***.
  - Causal inference
  - Descriptive epidemiology
  - Predictive Epidemiology
- (eg.) In the beginning of COVID-19 pandemic, measures to compare COVID-19 death rates among patients placed on ventilators vs those not on ventilators were needed. Rather than answering causal question, assessing the emerging public health crisis and identifying potential risk factors for severe clinical course were important. After understanding such fundamental nature, the interest moves to risk prediction, and RCT of vaccines.
  (cf.) https://www.nejm.org/doi/full/10.1056/NEJMp2002125
  (For Japanese) https://minato.sip21c.org/2019-nCoV-im3r.html#STUDYPERSPECTIVE)
- Simplest measure of association compares disease occurrence between exposed and unexposed. Risks or incidence rates can be compared by taking difference or ratio.

Effect measures

- To achieve a valid substitution for the counterfactual experiences, several designs (incl. crossover study, randomized experiment, choosing unexposed subjects who have the same or similar risk-factor profiles for disease as the exposed subjects) are possible.
- If we can assume the comparability, the effect of exposure can be measured by the following manners.
  - Absolute differences in incidence proportions and incidence rates
    - RD (risk difference) = attributable risk
      Risk (exposed) – Risk (unexposed)
    - IRD (incidence rate difference) = attributable rate
      IR (exposed) – IR (unexposed)
  - Relative risk
    - Relative effect = (RD)/(Risk in unexposed) = RR – 1
    - RR (risk ratio) = Risk (exposed) / Risk (unexposed)
    - IRR (incidence rate ratio) = IR (exposed) / IR (unexposed)

(Column) The Odds Ratio

- As the 3rd relative measure of association, the odds ratio (OR) is frequently used in epidemiology, ranges from 0 to infinity
  - Prevalence Odds = P/(1-P), as already explained, obtained from cross-sectional study
  - Incidence Odds (or Risk Odds) = A/B, in cohort study
    - A: cases occur among N people followed
    - B: the number who did not get disease = (N-A)
  - (In case-control study) Exposure Odds = (Number with past exposure)/(Number without past exposure), in cases and controls
- OR = (Odds in Exposed) / (Odds in Unexposed) = $(A_1/B_1)/(A_0/B_0)$
  - In cross-sectional study, it's measured as (Odds with attribute) / (Odds without attribute), similarly
  - In case-control study, it's measured as (Odds in Cases)/(Odds in Controls), which can be used to estimate the RR or IRR (Chapter 5)
  - In cohort studies, risk or incidence rate is directly available, so that using OR is not meaningful. The reason why OR is frequently reported comes from widely used logistic regression model (later Chapters)

Table 4.5. Comparison of absolute and relative effect measures

Measure	Numeric range	Dimensionality
Risk difference	[-1, +1]	None
Risk ratio	[0, ∞)	None
Incidence rate difference	(-∞, +∞)	1/time
Incidence rate ratio	[0, ∞)	None

Note: ∞ is actually division by zero when no disease occurred in unexposed group.

Relation between RR and IRR

- If risk remains less than about 0.20, for short time periods, RR = R(E)/R(U) = {IR(E) x time}/{IR(U) x time} = IR(E)/IR(U) = IRR
- For longer time periods, RR becomes different from IRR.
  - In the case of table 4-6, maximum possible R(E) cannot exceed 1, when R(U) is 0.44, RR must be less than 1/0.44 = 2.3 (1.96 << 2.3).
  - IR has no such restraint. In table 4-6, we can back calculate IR(E) and IR(U) from risk-data.
    - $14 \times (1 - IR(E))^{10} = 2 \rightarrow 1 - (2/14)^{0.1} = 0.1768\ldots = IR(E)$
    - $16 \times (1 - IR(U))^{10} = 9 \rightarrow 1 - (9/16)^{0.1} = 0.0559\ldots = IR(U)$
    - IRR = IR(E)/IR(U) = 3.16… (whereas the text says 3.4)
- For very shorter time periods, RR shrinks along with the length of the time interval: myocardial infarction risk in the next 10 seconds is almost zero. In such sense, IRR can be referred as instantaneous risk ratio. Both RR and IRR can be called as relative risk.

CAUSAL EFFECT MEASURES

- Different from court, assigning causation in a single person is impossible in epidemiology
- Epidemiology evaluates the proposition that the exposure is a cause of the disease in a theoretical sense, not infers what happened to any given person
- The exposed person can develop disease even if no causal connection exists between that exposure and disease
  - We cannot use the incidence proportion or incidence rates among exposed to measure causal effect
    - An observation “an infant receiving a vaccine develops autism” doesn’t mean that the vaccination caused autism
    - We have to contrast the **experience of exposed** with **what would happened to them if they were not exposed**

Counterfactual Ideal

- To measure the effect of exposure, ideally, it's necessary to compare the disease occurrence of exposed person with the hypothetical disease occurrence if the person were not exposed in the same time period (counterfactual ideal).
- It's impossible.
  - We don't have time-machine.
  - We cannot observe people in parallel universe (though it's written in Science-Fiction or Fantasy novels).
- Thus exposed group and unexposed group with similar background are compared. To control **the timing of exposure**, experimentally, crossover study is possible only if the exposure has a brief effect.
  - Compare A (exposed – washout – unexposed) and B (unexposed – washout – exposed)
  - The time sequence of exposure may affect the result, so that crossover study **also differs from counterfactual ideal**.

Measures of Effect (if comparability was achieved)

- RD and IRD provide direct measures of the absolute effect of an exposure
  - RD was referred as attributable risk in older textbook
  - IRD was referred as attributable rates in older textbook
- RR and IRR are effect measures on a relative scale
- Absolute and relative effect measures provide different meaning
  - The impact of an exposure on the health of population should be assessed by absolute effect. For most issues in public health, absolute measure may be useful
  - The extent to which disease among the exposed population is a consequence of exposure can be assessed by relative effect.
    - If IRR is 10 for extremely rare disease, the 10-fold increase of disease implies that the exposure accounts for almost all the disease among exposed, though it's still rare in the population
- In case-control studies (see Chapter 5), only relative effects are directly observable, but those can be converted into absolute measures by taking into account the overall rate or risk of disease occurrence in a population

(Column) When Risk Does Not Mean Risk

- Some people use the term risk as the mean of effect
  - RRs for lung cancer from asbestos exposure, 5 for young, 2.5 for older adults, some people wrongly say **“The risk of lung cancer from asbestos exposure is not as great among older people as among younger people”**.
    - RD between those exposed and those unexposed to asbestos is sure to be greater among older adults than younger.
    - Risk, or cumulative incidence of lung cancer itself steeply rise with age
  - If they know epidemiology, they must say **“The risk ratio of lung cancer from asbestos exposure is not as great among older people as among younger people”** or **“The effect of asbestos exposure on lung cancer is not as great among older people as among younger people”**
    - The latter expression is not wrong, because the term effect includes both relative effect and absolute effect, here it was used as relative

Examples

Table 4.6. Diarrhea during 10-day period in breastfed infants by antibody titer level

	Low	High	Total
Diarrhea	12	7	19
No diarrhea	2	9	11
Total	14	16	30
Risk	0.86	0.44	0.63

- 0.86 = 12/14
- RD = 0.86 – 0.44 = 0.42
- RR = 0.86/0.44 = 1.96
- The relative effect can be given by RD/(Risk in unexposed) = RR – 1 = 0.96 (it means 96% increase)

Table 4.7. Breast cancer cases and person-years of observation for women with TB, repeatedly exposed to multiple X-ray and unexposed

	Exposed	Unexposed	Total
BC cases	41	15	56
Person-yr	28010	19017	47027
Rate (/10000 yr)	14.6	7.9	11.9

- 14.6 = 41/28010x10000
- IRD = 14.6/10000 (yr^{-1}) – 7.9/10000 (yr^{-1}) = 6.7/10000 (yr^{-1})
- IRR = 14.6/7.9 = 1.86
- The relative effect can be given by IRR – 1 = 0.86 (86% greater rate of breast cancer among women exposed to the radiation)

(Column) Rounding: How many digits should be reported?

- In some published papers, RR is reported as 4.1, in others 4.0846
  - Basically the number of digits mean precision.
  - 4.0846 means that the true RR lies between 4.084 and 4.085, but very large study can produce such a high precision
- General rule is difficult to give
  - If the rule is reporting the first decimal, RR 4.1 is OK. However, if RR is less than 1, RR may be 0.7 or 0.8, which cause large rounding error (precision is lost too much)
  - If the rule is “using a constant number of meaningful digits”? That’s better than previous one, but in this rule, 0.98 must be distinguished from 0.99 but 1.00 may not be distinguished from 1.01 (If 2 digits are allowed, next to 1.0 is 1.1).
- Writers must use good judgement.
  - Values used in intermediate calculation should never be rounded
  - Rounding RR 1.41 to 1.4 may not be a large error, but rounding 1.25 to 1.2 or 1.3 causes 20% rounding error (Note: RR=1 means no effect)
  - Rounding a number ending 5 is customary to round upward (as in elementary school education), it causes upward bias. Instead, rounding to the nearest even number is better to avoid such bias (as in JIS rounding). Rounding 1.75 and 1.85 is both 1.8

Attributable fraction

Table 4.8. 1 yr disease risk for E and U			
	U	E	Total
Disease	900	500	1400
No disease	89100	9500	98600
Total	90000	10000	100000
Risk	0.01	0.05	0.014

Table 4.9. 1 yr disease risk for 3 level exposure				
	None	Low E	High E	Total
D	100	1200	1200	2500
No D	9900	58800	28800	97500
Total	10000	60000	30000	100000
Risk	0.01	0.02	0.04	0.025
RR	1	2	4	
Prop. in cases	0.04	0.48	0.48	

- RD = R(E) – R(U)
- Attributable fraction (AF) = RD/R(E) = {R(E) – R(U)}/R(E) = 1 – 1/RR
- The proportion of the disease burden among exposed people that is caused by the exposure
- In Table 4.8, AF=(0.05-0.01)/0.05=4/5=0.8 (80%)
- Among 1400 cases, 500 were exposed, the proportion was 500/1400 = 0.357. Overall AF for the population is 0.357x0.8=0.286
- Otherwise, 500x0.8=400 are attributable to exposure. 400/1400=0.286
- Total AF = $\sum (AF_i \times P_i)$
- In Table 4.9, total AF = 0 + {(2-1)/2}x0.48 + {(4-1)/4}x0.48 = 0.24+0.36 = 0.60
- Otherwise, {0 + 1200x(2-1)/2 + 1200x(4-1)/4}/2500 = 0.6