Berkeley School of Public Health

Age, period, and cohort effects

PHW250B

Time can affect measures of disease through:

- **Age effects:** Change in the measure of disease according to age, irrespective of birth cohort and calendar time

- **Cohort effects:** Change in the measure of disease according to year of birth, irrespective of age and calendar time
  - Can be thought of as an interaction between age and calendar time

- **Period effects:** Change in the measure of disease affecting an entire population at some point in time, irrespective of age and birth cohort

Time can affect measures of disease through:

- **Age effects:** For almost all diseases, a person's age is associated with their risk

- **Cohort effects:** Are not necessarily only related to the time of birth. For example, individuals in different birth cohorts may have different diets that pose different health risks over their lifetimes.

- **Period effects:** Can be caused by introduction of new medication or preventive interventions, historical events (e.g., nuclear bombing)

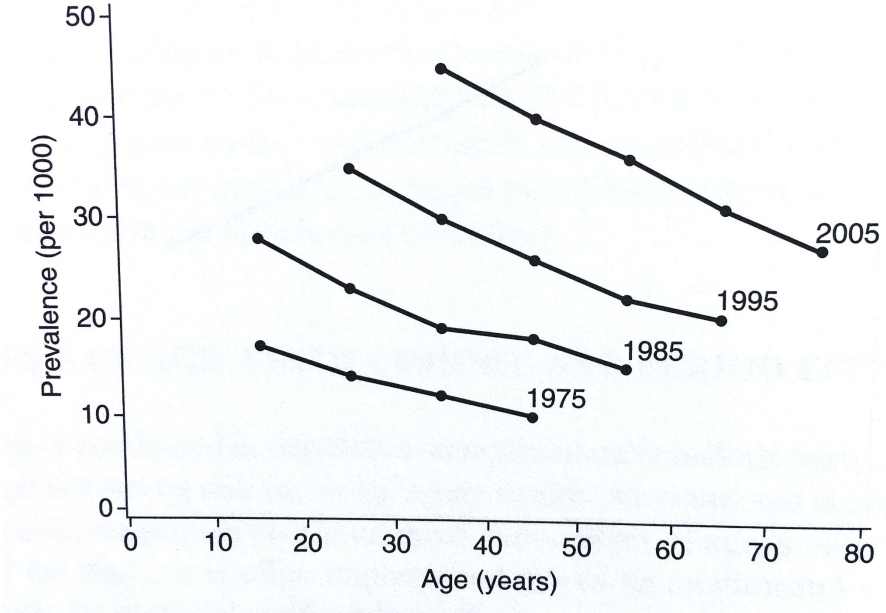
Hypothetical data that underlies the graphs on slides 3-4

- This table presents data from the same population in which cross-sectional studies were conducted in 1975, 1985, 1995, and 2005.

Age Group (Years)	*Midpoint (Years)*	*Survey Date*			
		1975	*1985*	*1995*	*2005*
		Prevalence (per 1000)			
10–19	15	17	28		
20–29	25	14	23	35	
30–39	35	12	19	30	45
40–49	45	10	18	26	40
50–59	55		15	22	36
60–69	65			20	31
70–79	75				27

Szklo & Nieto, 2nd ed.

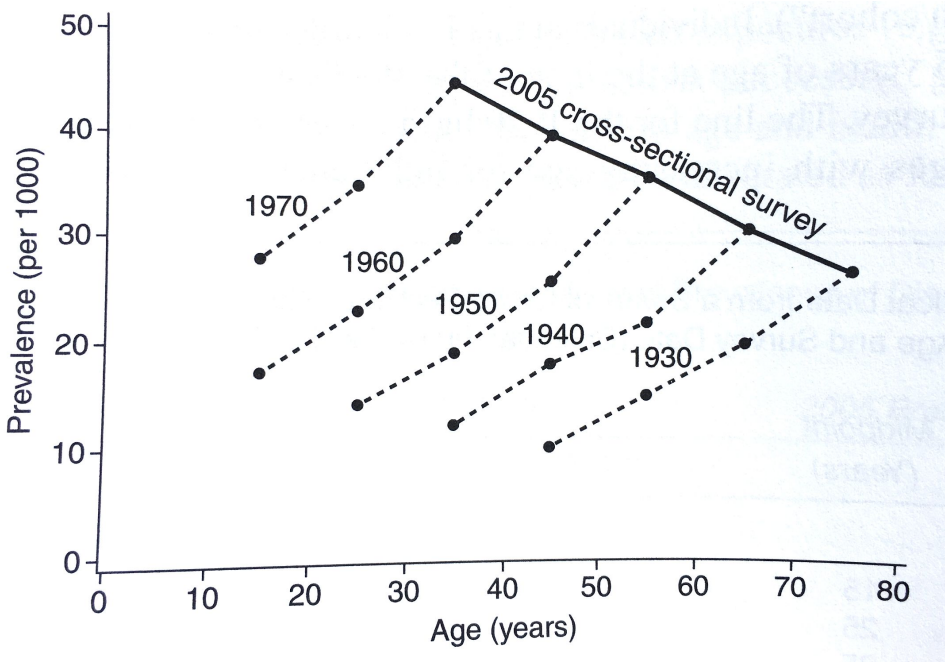
Example of age effects



- Same data points, in the next three plots, different lines connecting the dots.
- Lines in this plot connect data points from each survey.
- The prevalence decreases with age for each survey conducted in 1975, 1985, 1995, 2005

Szklo & Nieto, 2nd ed.

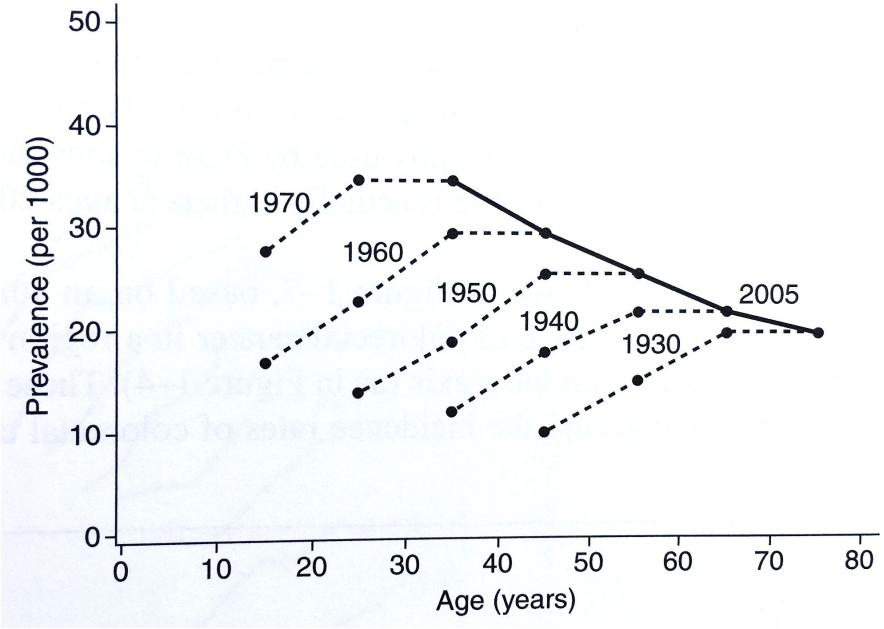
Example of cohort effects



- Lines in this plot connect data points for different birth cohorts.
- The solid line represents the observed cross-sectional pattern in the 2005 survey.
- The prevalence *decreases* with age for each survey conducted in 1975, 1985, 1995, 2005.
- There is a strong cohort effect: the prevalence is strongly affected by the person's year of birth.
- The prevalence is higher in younger vs. older cohorts.
- The fact that more recent cohorts have higher rates overwhelms the increase in prevalence associated with age.

Szklo & Nieto, 2nd ed.

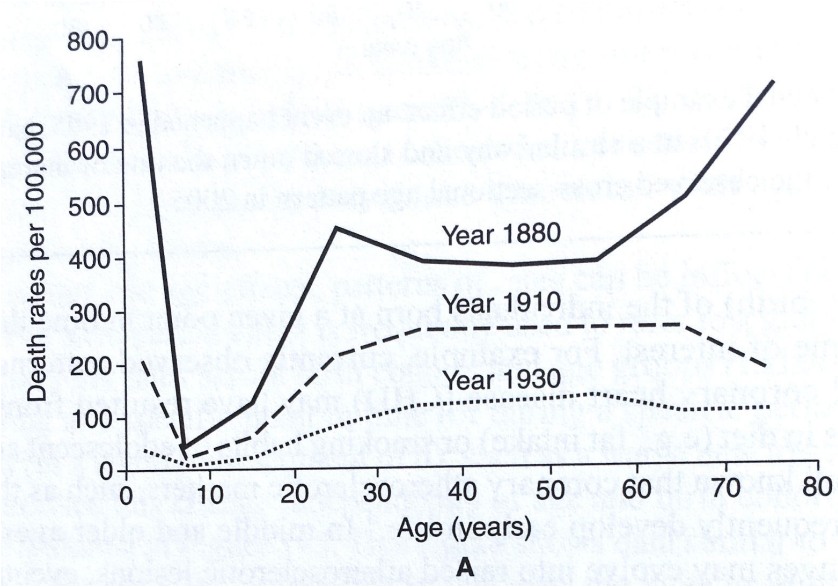
Example of period effects



- Plot with different underlying data
- Dashed lines in this plot connect data points for different birth cohorts
- The solid line represents the observed cross-sectional pattern in the 2005 survey.
- An event happened in 1995 that affected birth cohorts from 1930-1970

Szklo & Nieto, 2nd ed.

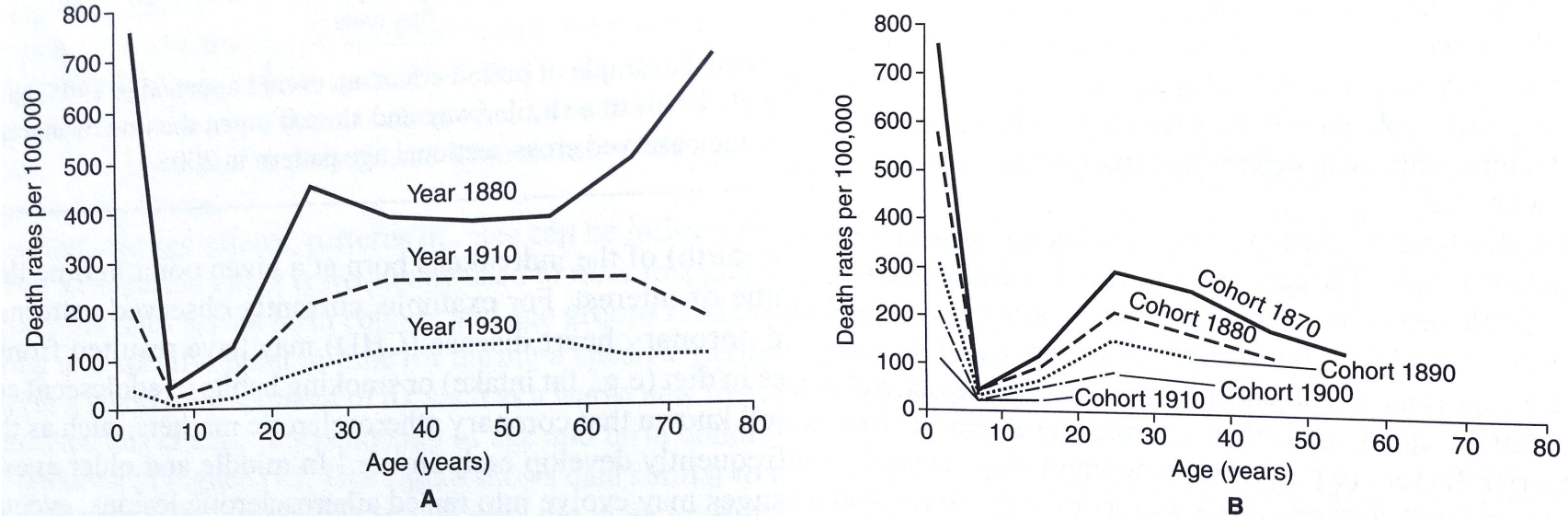
Real example: Frost's study of tuberculosis



- Wade Hampton Frost studied how tuberculosis mortality varied by age.
- "Looking at the 1930 curve, the impression given is that nowadays an individual encounters his greatest risk of death from tuberculosis between the ages of 5o and 60. But this is not really so…"

Szklo & Nieto, 2nd ed.

Real example: Frost's study of tuberculosis



- "... the people making up the 1930 age group 30 to 60 have, in earlier life, passed through *greater* mortality risk."

Szklo & Nieto, 2nd ed.

Summary of key points

- Age, period, and cohort effects are common in most epidemiologic studies.
- As epidemiologists, we need to think carefully about how we present data (e.g., which lines connect the individual data points?) because they can reveal different patterns.
- The type of effect that is present may have different implications for intervention.
  - A strong age effect may suggest targeting a specific age group for intervention
  - A strong cohort effect may suggest targeting a specific cohort for intervention
  - A strong period effect may explain the positive or negative impact of a policy or historical event and influence future policymaking

Berkeley School of Public Health

Types of populations

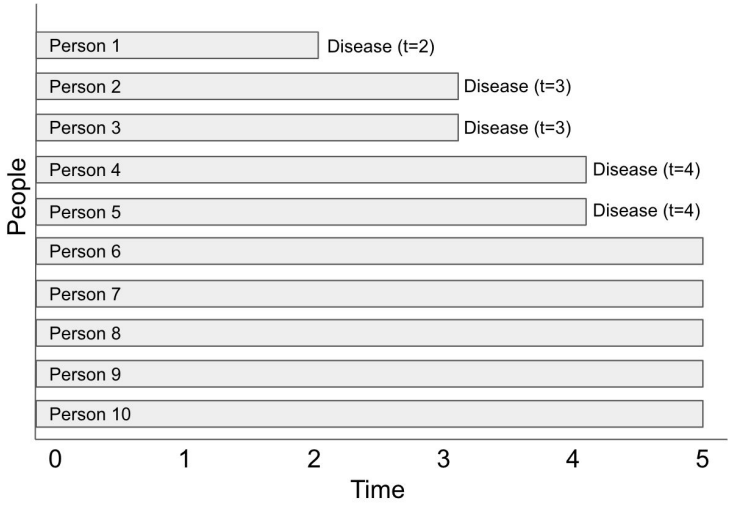
PHW250B

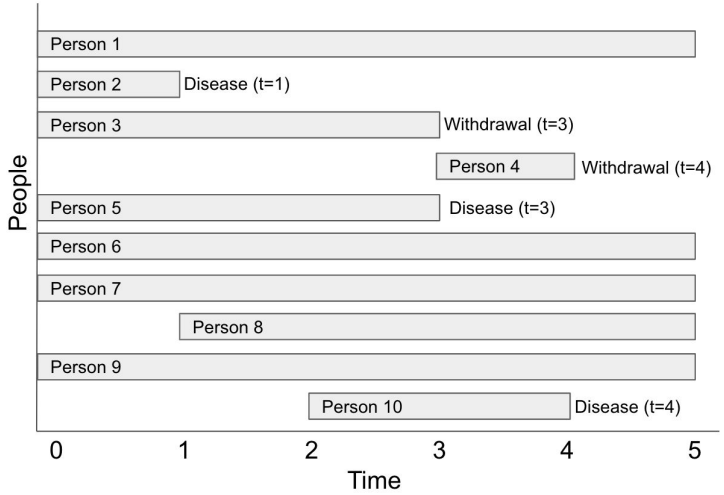
Closed vs. open populations

Closed population: a population in which there are no entries during the follow-up period and no losses to follow-up. *(sometimes called "fixed" population)*

Open population: a population in which the person time experience can accrue from a changing roster of individuals. Individuals can enter during the follow-up period. *(sometimes called "dynamic" population)*

(most common in practice)





Closed vs. open populations

- How do people enter open populations?
  - Birth
  - Migration in
- How do people exit open populations?
  - Death
  - Migration out
  - Termination of the study
  - Use of certain medical procedures that change a person's status to ineligible for the study
    - E.g., a person who has a hysterectomy is no longer eligible for a study of uterine cancer
- People may exit and re-enter open populations

Closed vs. open populations

- Whether a population is open or closed depends on the time scale used to define a population
- Example: All people who ever used a specific drug
  - If the time scale is the time when a person started using the drug to when they stopped using a drug, then this is a **closed population**.
  - If the time scale is the calendar time, this is an **open populatio**n because new drug users may be added to the population over time.

Steady state

- A **steady state** population is one in which the number of people entering and existing is balanced within a period of time across age, sex, and other factors that affect disease risk.
- This property is only relevant to **open populations**.

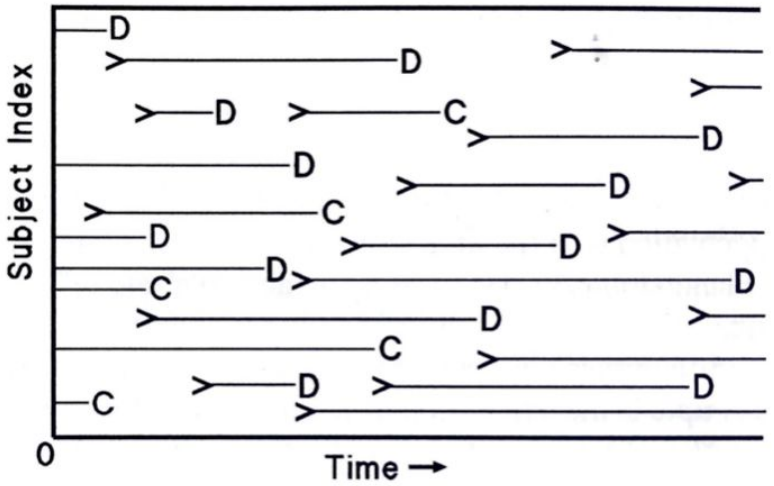


FIGURE 3–3 • Composition of an open population in approximate steady state, by time; > indicates entry into the population, D indicates disease onset, and C indicates exit from the population without disease.

Rothman ME3

Summary of key points

- Closed population
- Open population
  - Steady state populations are a type of open population
- These population classifications will affect the choice of method used to calculate incidence.

Berkeley School of Public Health

Calculating incidence density and incidence rates

PHW250B

Review: incidence rates

- **Numerator**: number of incident (new) cases of disease
- **Denominator**: person-time at risk during follow-up period
- **Units:** time^{-1}
- **Range:** 0 to infinity
- **Interpretation**: the rate at which disease events occur in the population at risk at any given point in time
- The instantaneous rate for each individual cannot be calculated directly, but we can average incidence rates over a period of time and use that as a proxy for an individual's rate
- **Synonyms:** Epidemiologists use the terms "incidence density" and "incidence rate" interchangeably.
  - Szklo & Nieto use "incidence density" for incidence calculated from individual level data and "incidence rates" for incidence calculated from aggregate data.

Review: incidence rates cont.

- Incidence rates can be calculated for closed or open populations.
- Study participants can be followed for different amounts of time.
- It does not distinguish between people who never developed the disease or just did not develop the disease during the time of the study

How to choose the unit of time for incidence rates

- Incidence rates can be expressed with different units of time
- The following are equivalent:
  - 0.024 per person-day
  - 2.4 per 100 person-days
  - 8.76 per person-years
- Person-years are commonly used when the disease is rare
- Other units may be more appropriate for more common diseases

Examples of different time units for incidence rates

TABLE 2-5 Examples of person-time units according to the frequency of events under investigation.

Population	Event studied	Person-time unit typically used
General	Incident breast cancer	Person-years
General	Incident myocardial infarction	Person-years
Malnourished children	Incident diarrhea	Person-months
Pancreatic cancer cases	Death	Person-months
Influenza epidemic	Incident influenza	Person-weeks
Infants with acute diarrhea	Recovery	Person-days

Szklo 4th ed.

How to calculate person-time

N' : population at risk
Δt_i : duration of follow-up for person i

- If you have **individual level** data and each person's exact time contribution is known:

$$PT = \sum_{i=1}^{N'} \Delta t_i$$

- If you have **aggregated** data:
  - Assumes the population is in steady state

$$PT = N'(\Delta t)$$

How to calculate incidence rates

- If you have **individual level** data
  - e.g., in a cohort study
  - Incidence rate = number of events / total person-time
- If you have **aggregated** data
  - e.g., using census/ surveillance data
  - Incidence rate = number of events / average population during follow-up period
  - Use of average population assumes a constant incidence rate during the follow-up period

Incidence rate based on individual level data

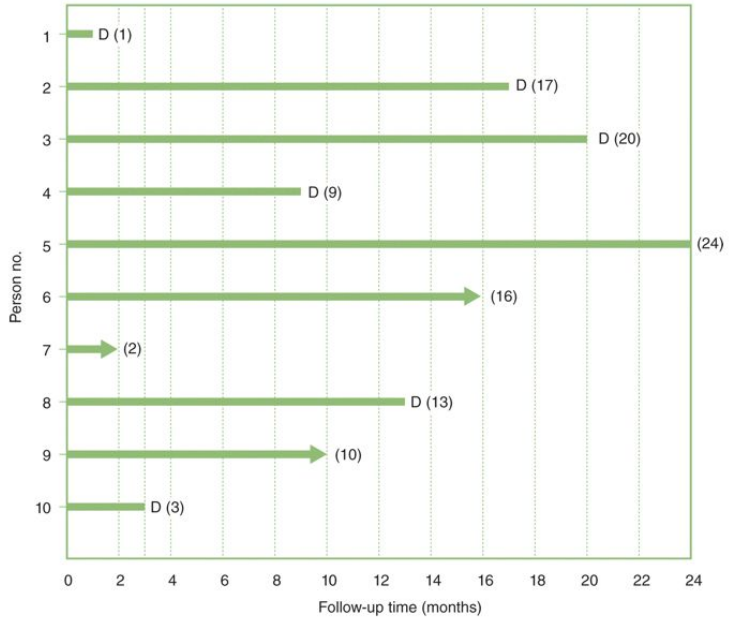


FIGURE 2-2 Same cohort as in Figure 2-1, with person-time represented according to time since the beginning of the study. D, death; arrow, censored observation; ( ), duration of follow-up in months (all assumed to be exact whole numbers).

Total follow-up (in months)	Total person years
1	0.083
17	1.417
20	1.667
9	0.750
24	2.000
16	1.333
2	0.167
13	1.083
10	0.833
3	0.250
115 months	9.583 years

Incidence rate = number of events / total person-time

= 6 / 9.583 person-years
= 0.63 per person-year

Szklo 4th ed.

Incidence rate based on aggregated data

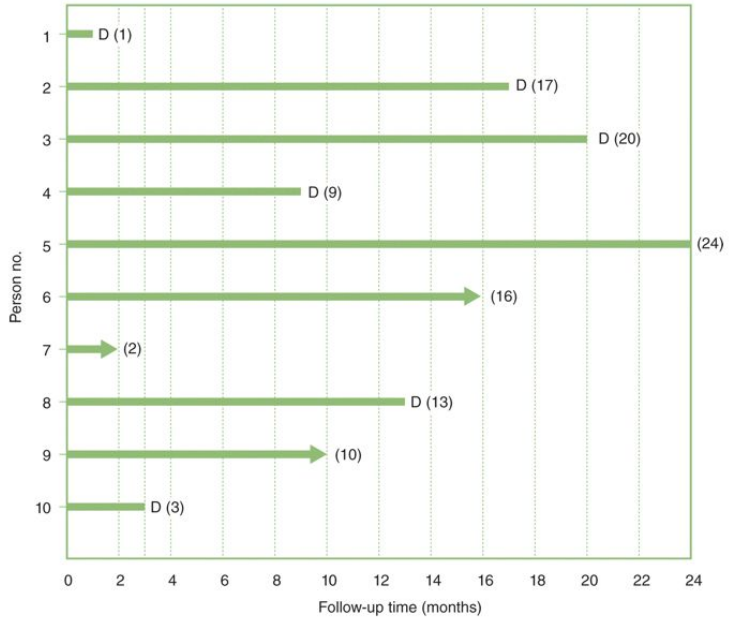


FIGURE 2-2 Same cohort as in Figure 2-1, with person-time represented according to time since the beginning of the study. D, death; arrow, censored observation; ( ), duration of follow-up in months (all assumed to be exact whole numbers).

- Incidence rate = Number of events / average population
- **Average population:** average of population at beginning of follow-up and at the end of follow-up
- Average population = (10 + 1)/2 = 5.5
- Incidence rate =
  - 6 / 5.5 = 1.09 per person over 2 years
  - 0.545 per person-year

Szklo 4th ed.

Does incidence calculated from individual and aggregate data provide the same estimate?

- Yes, if withdrawals, additions to the population (e.g., births or migration in), and disease events occur uniformly over time

$$\text{Incidence rate using aggregate data} = \frac{\#\text{ events}}{\frac{\text{average population (n)}}{\text{time (t)}}} = \frac{\#\text{ events}}{n \times t} = \text{Incidence rate using individual data}$$

This is convenient for studies that must rely on aggregate data. For example, in occupational epidemiology, it is common to estimate age-specific rates using aggregate vital statistics data since individual level data is not always available.

Assumptions when calculating incidence rates

1. **Independence of censoring and survival**
2. **Lack of secular trends**
3. **Risk of the event remains approximately constant** over time during the interval of interest

Assumption 1: Independence of censoring and survival (review from cumulative incidence video)

- Censored individuals have the same probability of the event after censoring as those remaining under observation
  - i.e., censoring is independent of survival
- Example: if patients withdraw from a study because they are sicker than those who do not withdraw, over time, the remaining study population would have patients with decreasing risk of illness, causing incidence to be underestimated.
- This assumption is difficult to make when the disease shares strong risk factors with diseases associated with mortality.

Assumption 2: lack of secular trends

(review from cumulative incidence video)

- There are no secular trends in individual characteristics, exposures, or interventions during follow-up that affect the disease.
- Birth cohort and period effects can produce secular trends that bias incidence rate estimates.
- Example: It would not be appropriate to estimate survival from diagnosis of all patients with insulin-dependent diabetes from 1915 through 1935 because this group would include:
  - Patients diagnosed before the introduction of insulin, who had a much lower chance of survival
  - Patients diagnosed after the introduction of insulin, who had a much higher chance of survival
  - It would be more appropriate to calculate incidence rates separately for those time periods

Assumption 3: Risk of the event remains approximately constant over time during the interval of interest

- The risk of an individual living five units of time within the interval is equivalent to that of five individuals living one unit each
- This assumption is not always valid. For example, in studies of smoking, the risk of bronchitis for 1 smoker followed for 30 years is not likely to be the same as that of 30 smokers followed for 1 year because of the cumulative effect of smoking.
- To weaken this assumption, you can calculate incidence within shorter time intervals

Depiction of the assumption of constant risk over the follow-up period

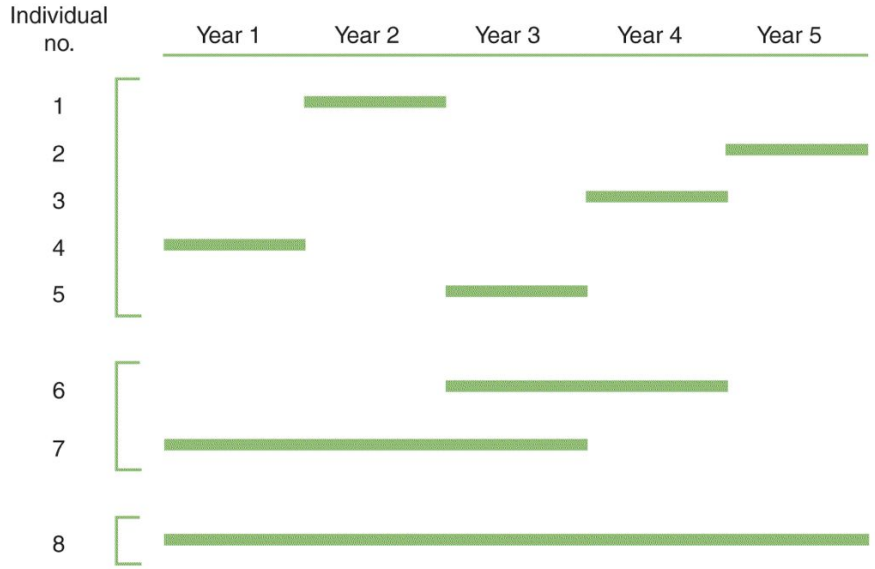


FIGURE 2-7 Follow-up time for eight individuals in a hypothetical study. It is assumed that the sum of the person-time units for individuals nos. 1 to 5 (with a short follow-up time of 1 year each) is equivalent to the sum for individuals nos. 6 and 7 (with follow-up times of 2 and 3 years, respectively) and to the total time for individual no. 8 (who has the longest follow-up time, 5 years). For each group of individuals (nos. 1–5, 6 and 7, and 8), the total number of person-years of observation is 5.

Szklo 4th ed.

Summary of key points

- Incidence rates capture the pace at which disease events occur in the population at risk.
- You can calculate incidence rates using either individual-level or aggregated data.
- Incidence rates can be calculated for closed or open populations.
- Study participants can be followed for different amounts of time.
- It is important to assess the assumptions made when calculating incidence rates to determine whether they are appropriate in a given study setting. When assumptions are violated, incidence rates will be biased.

Berkeley School of Public Health

Calculating cumulative incidence - Part 1

PHW250 F - Jade Benjamin-Chung

Outline

- Recap: cumulative incidence and risk
- Choosing the appropriate approach to calculating cumulative incidence
  - Simple cumulative method
  - Actuarial Method
  - Kaplan-Meier Method
  - Density Method

Review: cumulative incidence

- **Numerator**: number of incident (new) cases of disease
- **Denominator**: population at risk during follow-up period
- **Units:** unitless
- **Range:** 0 to 1
- **Interpretation**: It is the proportion of a closed population at risk that becomes diseased within a given period of time.
- Rothman distinguishes between:
    - The incidence proportion for a specific interval of time and
    - The cumulative incidence as the sum of incident rate times the duration of each interval over multiple intervals.
  - Confusingly, most epidemiologists use these terms interchangeably.
- In this class we will generally use the term “cumulative incidence”

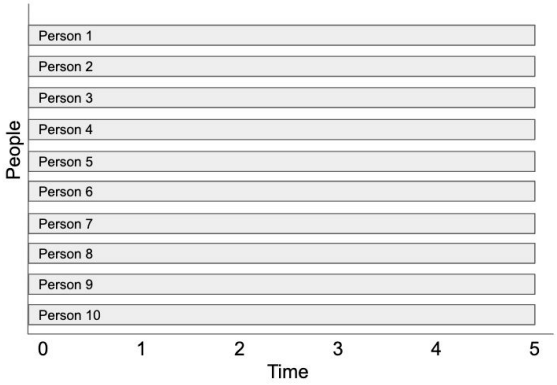
Risk vs. cumulative incidence

- **Risk:** the probability that a disease-free **individual** develops a disease within a specific time period, conditional on that individual not dying from any other disease during the period

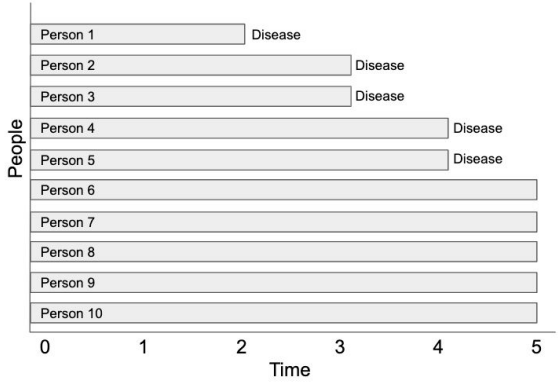
- **Cumulative incidence:** proportion of **subjects** who develop the disease during the observation period

When does the cumulative incidence equal the average risk?

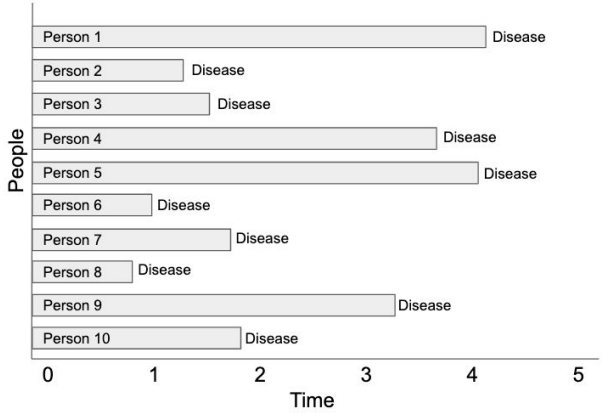
(1) All non-cases have the same length of follow-up



(2) There is no loss to follow-up / withdrawal



(3) All cohort members develop disease during follow-up



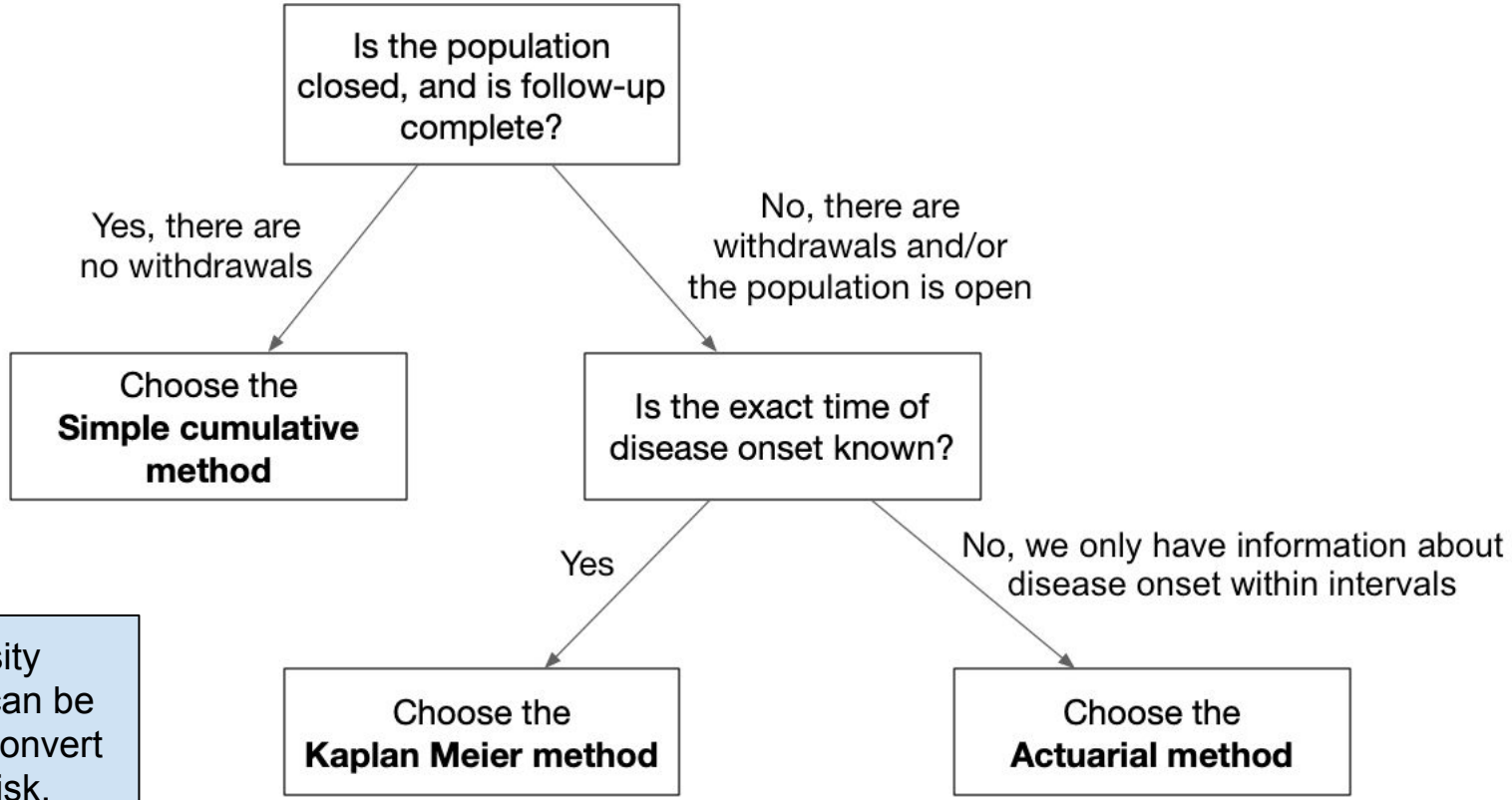
These scenarios occur infrequently in practice. **Thus, the cumulative incidence is almost always an estimate of the risk.**

Kleinbaum et al., *Epidemiologic Research;* 1982

Outline

- Recap: cumulative incidence and risk
- Choosing the appropriate approach to calculating cumulative incidence
  - Simple cumulative method
  - Actuarial Method
  - Kaplan-Meier Method
  - Density Method

Choosing a method to calculate cumulative incidence



The density method can be used to convert rates to risk.

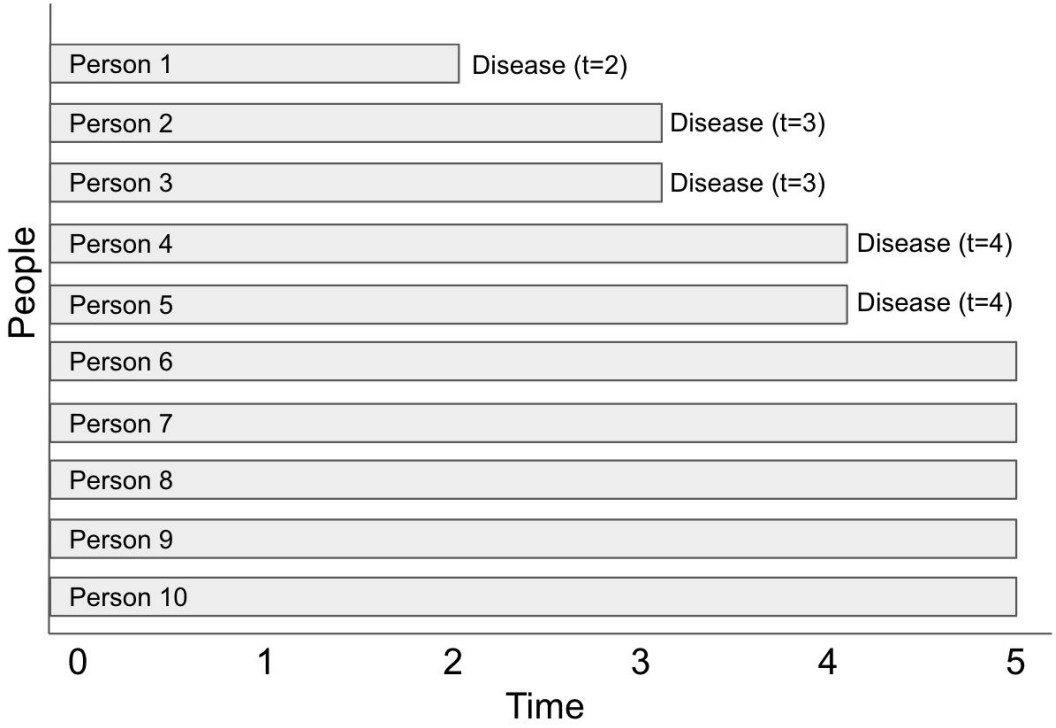
Outline

- Recap: cumulative incidence and risk
- Choosing the appropriate approach to calculating cumulative incidence
  - Simple cumulative method
  - Actuarial Method
  - Kaplan-Meier Method
  - Density Method

Simple cumulative method

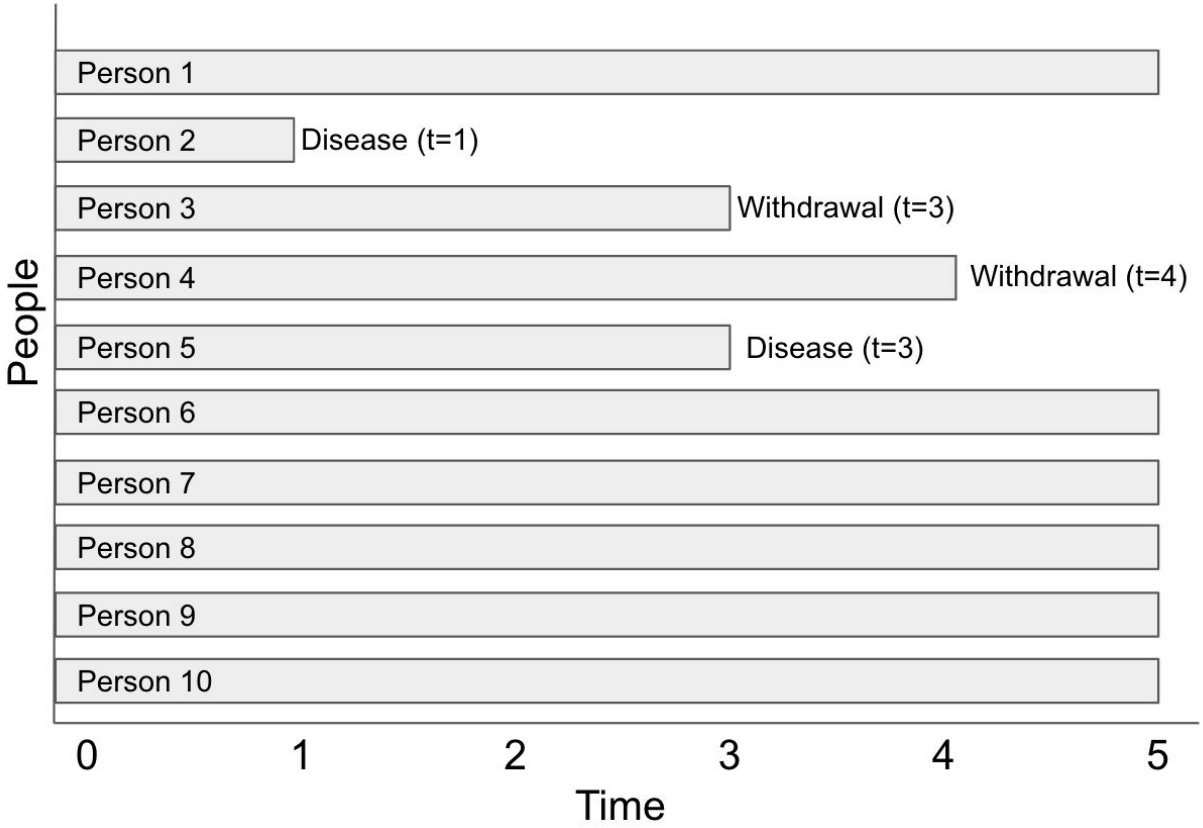
- $CI_{(t0, t)} = I / N'_0$
- I: number of incident cases within the follow-up period
- N'_0: number of people at risk at the beginning of follow-up
- Assumes there are no withdrawals, i.e., that all participants are followed for the entire follow-up period
- Appropriate for short time frames (e.g., food-borne illness outbreak)
- This measure is often called the “attack rate”
- If the population is closed and there is no attrition during follow-up, the cumulative incidence is equivalent to the risk.
  - $R_{(t0, t)} = CI_{(t0, t)} = I / N'_0$

Example 1: Simple cumulative method



- $CI_{(t0, t)} = I / N'_0$
- **I**: number of incident cases within the follow-up period
- N'_0: number of people at risk at the beginning of follow-up
- CI = 5 / 10 = 0.5

Example 2: The simple cumulative method is not appropriate because there are withdrawals!



We will use a different method instead.

The simple cumulative method would assume that person 3 and person 4 were followed up to time 5 and that they did not develop the disease.

However, we do not know their disease status after the time of withdrawal.

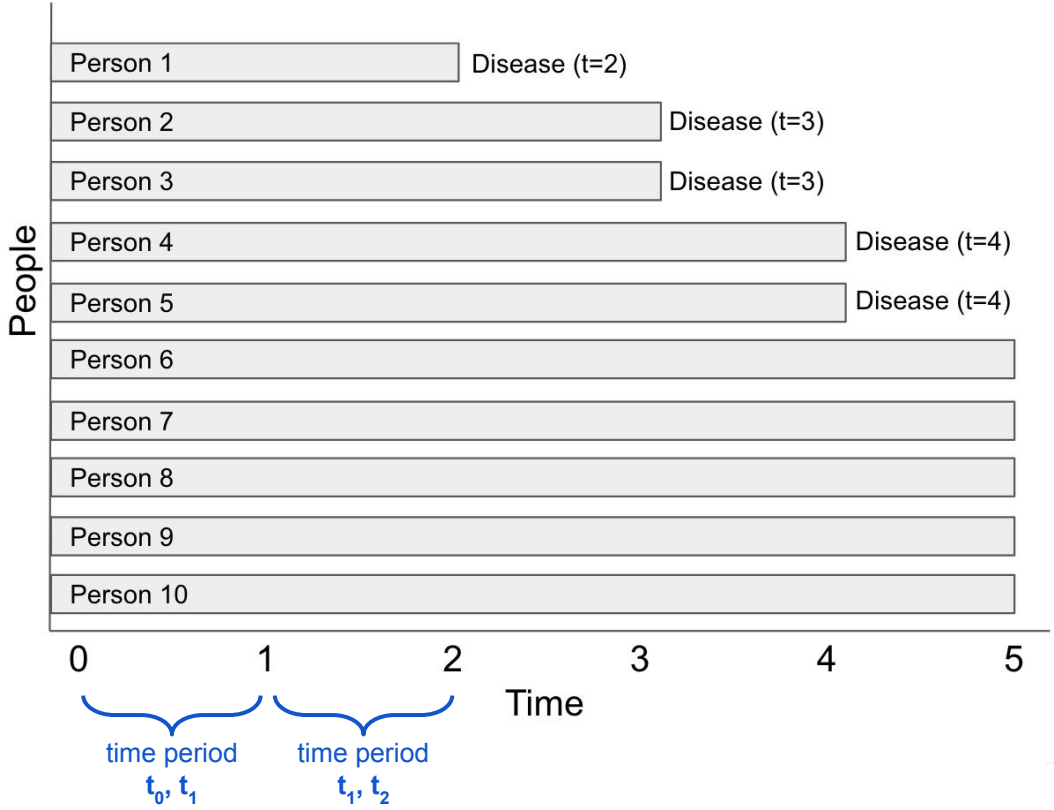
Outline

- Recap: cumulative incidence and risk
- Choosing the appropriate approach to calculating cumulative incidence
  - Simple cumulative method
  - Actuarial Method
  - Kaplan-Meier Method
  - Density Method

Actuarial Method

- Appropriate for incomplete follow-up
- Intuitively, this method involves:
  1. Break the follow-up time into small time intervals
  2. Treat the population as a closed cohort within that time interval (even if the study population was open)
  3. Estimate the **risk** in each interval assuming any withdrawals occurred halfway through the interval
     i. For this reason, interval length should be relatively short.
     ii. The interval risk is a conditional probability — it conditions on whether a person was at risk (alive and not censored) at the event time.
  4. Calculate the **cumulative incidence** that accumulated over all intervals

Terminology



j: indicates the interval

time period t_{j-1}, t_j is the time period spanning t_{j-1} to t_j

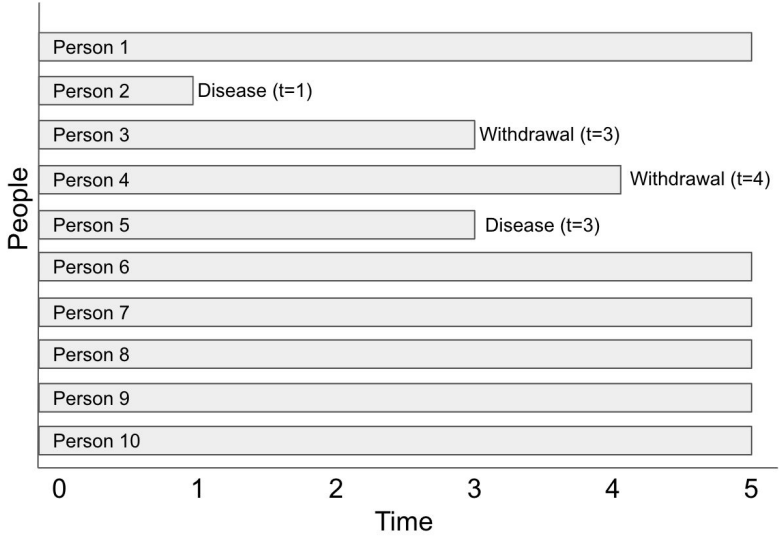
N'_{0j} : number of disease-free individuals at the beginning of interval **j**

In the figure to the left,
$N'_0 = 10$
$N'_{02} = 10$
$N'_{04} = 7$

Choice of interval length

- Choice of interval length is based on the extent to which incidence changes over time
  - The goal is for events and withdrawals to occur at an even rate throughout the interval
- Intervals do not have to be the same duration
- **Example:**
  - Study of survival after heart attack
  - Short intervals could be used soon after symptom onset when the probability of death is high and changes quickly
  - Longer intervals can be used later after symptom onset

Actuarial method

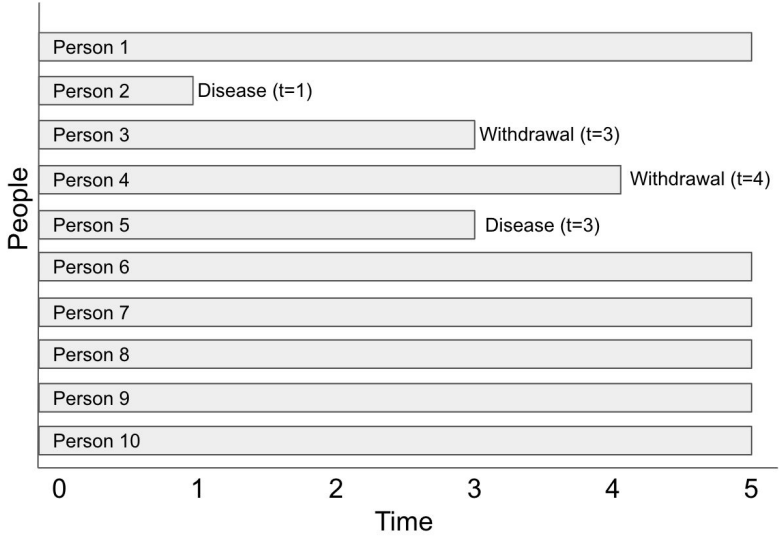


This is the same as Example 2 shown for the simple cumulative incidence.

We use this method when we measure disease in intervals. This means that we record disease onset for person 2 at time 1, but that person could have developed disease at any time between time 0 and time 1.

Actuarial method

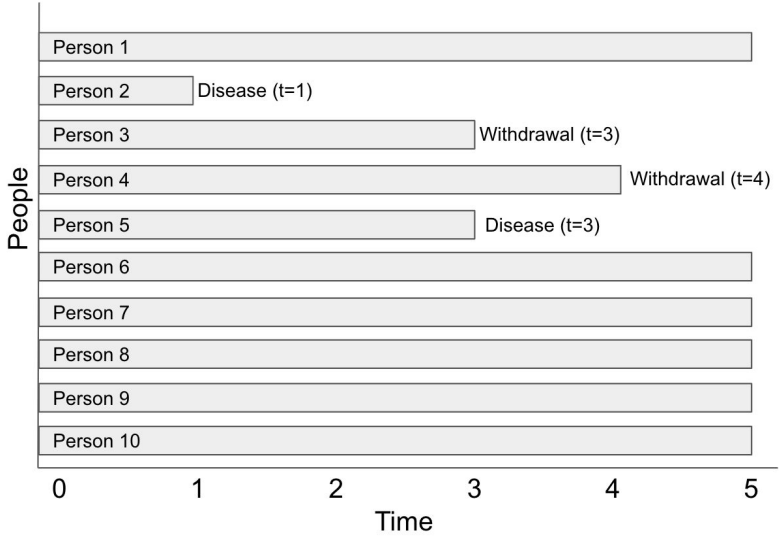
Step 1: Calculate the population at risk at the beginning of each interval (N'_{0j})



Interval j	time period t_{j-1}, t_j	pop. at risk N'_{0j}	incident disease I_j	Withdrawals W_j	interval risk (conditional risk) R_j	Interval survival S_j
1	0,1	10				
2	1,2	9				
3	2,3	9				
4	3,4	7				
5	4,5	6				

Actuarial method

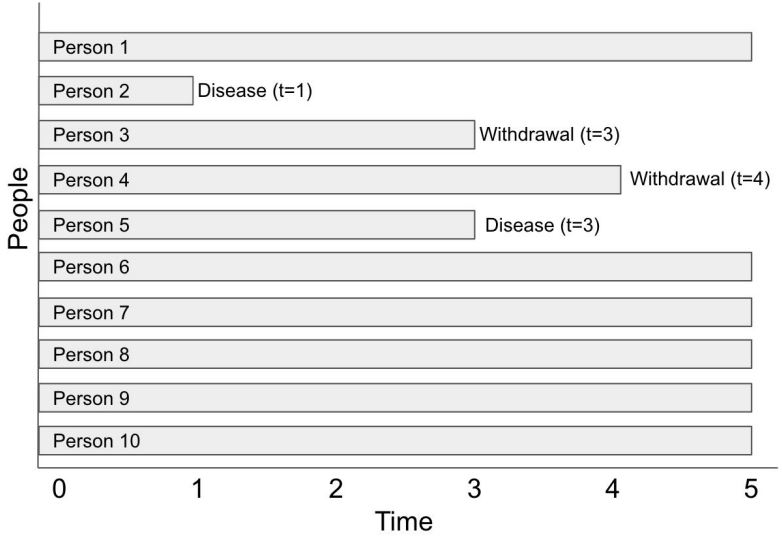
Step 2: Count the number of incident cases within each interval (I_j)



Interval j	time period t_{j-1}, t_j	pop. at risk N'_{0j}	incident disease I_j	Withdrawals W_j	interval risk (conditional risk) R_j	Interval survival S_j
1	0,1	10	1			
2	1,2	9	0			
3	2,3	9	1			
4	3,4	7	0			
5	4,5	6	0			

Actuarial method

Step 3: Count the number of withdrawals within each interval (W_j)

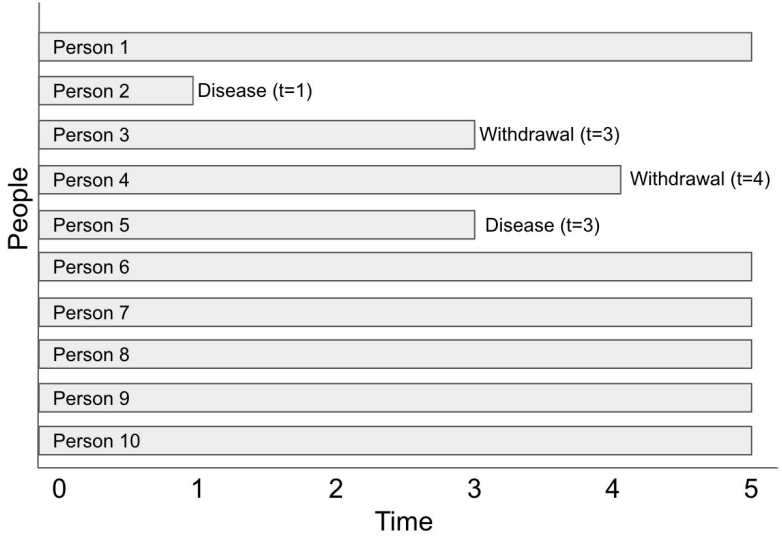


Interval j	time period t_{j-1}, t_j	pop. at risk N'_{0j}	incident disease I_j	Withdrawals W_j	interval risk (conditional risk) R_j	Interval survival S_j
1	0,1	10	1	0		
2	1,2	9	0	0		
3	2,3	9	1	1		
4	3,4	7	0	1		
5	4,5	6	0	0		

Actuarial method

Step 4: Calculate the risk within each interval (R_j) using the formula: $$R_j = \frac{I_j}{N'_{0j} - W_j/2}$$

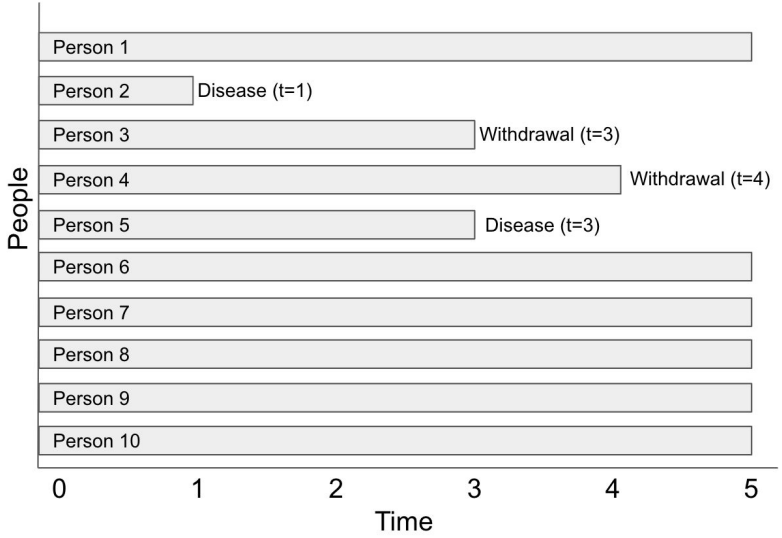
Keep 3 decimal points!!



Interval j	time period t_{j-1}, t_j	pop. at risk N'_{0j}	incident disease I_j	Withdrawals W_j	interval risk (conditional risk) R_j	Interval survival S_j
1	0,1	10	1	0	1 / 10 = 0.100	
2	1,2	9	0	0	0 / 9 = 0.000	
3	2,3	9	1	1	1 / (9 - 1/2) = 0.118	
4	3,4	7	0	1	0 / (7 - 1/2) = 0.000	
5	4,5	6	0	0	0 / 6 = 0.000	

Actuarial method

Step 5: Calculate the survival within each interval (S_j) using the formula: $S_j = 1 - R_j$

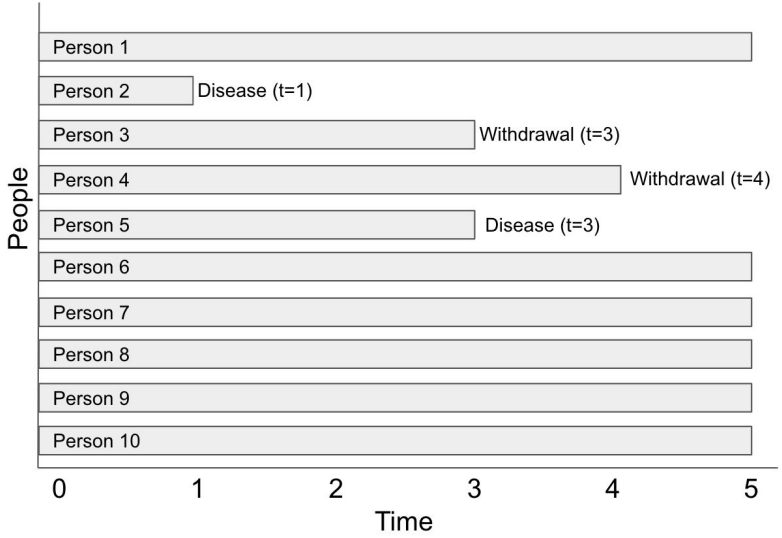


Interval j	time period t_{j-1}, t_j	pop. at risk N'_{0j}	incident disease I_j	Withdrawals W_j	interval risk (conditional risk) R_j	Interval survival S_j
1	0,1	10	1	0	1 / 10 = 0.100	
2	1,2	9	0	0	0 / 9 = 0.000	
3	2,3	9	1	1	1 / (9 - 1/2) = 0.118	
4	3,4	7	0	1	0 / (7 - 1/2) = 0.000	
5	4,5	6	0	0	0 / 6 = 0.000	

Actuarial method

Step 5: Calculate the survival within each interval (S_j) using the formula: $S_j = 1 - R_j$

Keep 3 decimal points!!

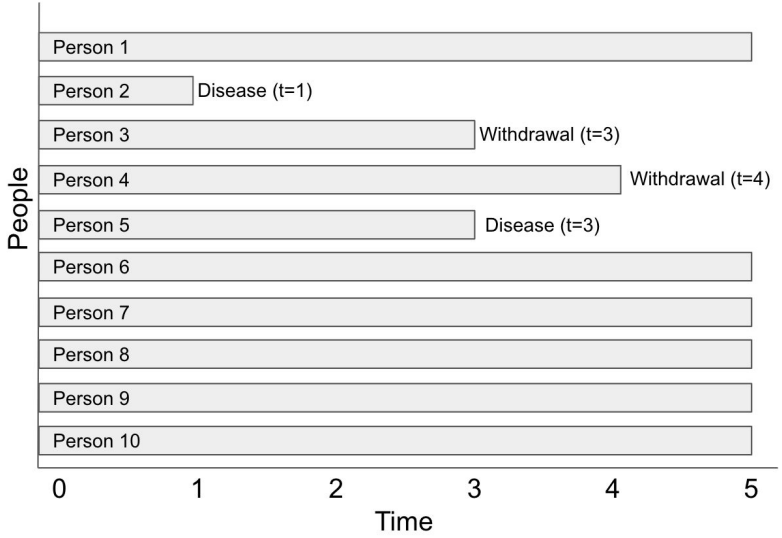


Interval j	time period t_{j-1}, t_j	pop. at risk N'_{0j}	incident disease I_j	Withdrawals W_j	interval risk (conditional risk) R_j	Interval survival S_j
1	0,1	10	1	0	1 / 10 = 0.100	0.900
2	1,2	9	0	0	0 / 9 = 0.000	1.000
3	2,3	9	1	1	1 / (9 - 1/2) = 0.118	0.882
4	3,4	7	0	1	0 / (7 - 1/2) = 0.000	1.000
5	4,5	6	0	0	0 / 6 = 0.000	1.000

Actuarial method

Step 6: Calculate the cumulative incidence for the entire follow-up period using the product limit formula: $CI_{(t_0,t_j)} = 1 - \prod_j (1 - R_j) = 1 - \prod_j S_j$

Keep 3 decimal points!! Then round at the end.



Interval j	time period t_{j-1}, t_j	pop. at risk N'_{0j}	incident disease I_j	Withdrawals W_j	interval risk (conditional risk) R_j	Interval survival S_j
1	0,1	10	1	0	1 / 10 = 0.100	0.900
2	1,2	9	0	0	0 / 9 = 0.000	1.000
3	2,3	9	1	1	1 / (9 - 1/2) = 0.118	0.882
4	3,4	7	0	1	0 / (7 - 1/2) = 0.000	1.000
5	4,5	6	0	0	0 / 6 = 0.000	1.000

$CI_{(t0,tj)}$ = 1 - (0.900 x 1.000 x 0.882 x 1.000 x 1.000) = 0.2062 (rounds to 0.206)

Actuarial method steps

1. Calculate the population at risk at the beginning of each interval (N'_{0j})
2. Count the number of incident cases within each interval (I_j)
3. Count the number of withdrawals within each interval (W_j)
4. Calculate the risk within each interval (R_j) using the formula

$$R_j = \frac{I_j}{N'_{0j} - W_j/2}$$

5. Calculate the survival within each interval (S_j) using the formula

$$S_j = 1 - R_j$$

6. Calculate the cumulative incidence for the entire follow-up period (t_0 to t_j) using the product limit formula $CI_{(t_0,t_j)} = 1 - \prod_j (1 - R_j) = 1 - \prod_j S_j$

The product-limit formula

Cumulative risk from t_0 to t_j : $CI_{(t_0,t_j)} = 1 - \prod_j (1 - R_j) = 1 - \prod_j S_j$

The product-limit formula

Cumulative risk from t_0 to t_j : $CI_{(t_0,t_j)} = 1 - \prod_j (1 - R_j) = 1 - \prod_j S_j$

$$S_j = \underbrace{\frac{N'_{0j} - I_j}{N'_{0j}}}_{1} = \underbrace{\frac{N'_{0j+1}}{N'_{0j}}}_{2}$$

1. The surviving population at time j is equal to the population at risk at the beginning of interval j minus the number of incident cases at time j divided by the population at risk at the beginning of interval j.

2. This is equal to the population at risk at the beginning of the next interval (j + 1) divided by the population at risk at the beginning of interval j.

The product-limit formula

Cumulative risk from t_0 to t_j : $CI_{(t_0,t_j)} = 1 - \prod_j (1 - R_j) = 1 - \prod_j S_j$

$$S_j = \boxed{1}\ \frac{N'_{0j} - I_j}{N'_{0j}} = \boxed{2}\ \frac{N'_{0j+1}}{N'_{0j}}$$

1. The surviving population at time j is equal to the population at risk at the beginning of interval j minus the number of incident cases at time j divided by the population at risk at the beginning of interval j.

2. This is equal to the population at risk at the beginning of the next interval (j + 1) divided by the population at risk at the beginning of interval j.

$$S_{(j=1,j=4)} = \frac{N'_{04}}{N'_{01}} = \frac{N'_{04}}{N'_{03}} \times \frac{N'_{03}}{N'_{02}} \times \frac{N'_{02}}{N'_{01}}$$

3. The proportion of the original population that remains at risk (i.e., that survives) at the end of a follow-up period with multiple intervals (e.g., at interval 4) is equal to the product of #2 above for each interval.

The product-limit formula

Cumulative risk from t_0 to t_j : $CI_{(t_0,t_j)} = 1 - \prod_j (1 - R_j) = 1 - \prod_j S_j$

$$S_j = \boxed{1}\ \frac{N'_{0j} - I_j}{N'_{0j}} = \boxed{2}\ \frac{N'_{0j+1}}{N'_{0j}}$$

1. The surviving population at time j is equal to the population at risk at the beginning of interval j minus the number of incident cases at time j divided by the population at risk at the beginning of interval j.

2. This is equal to the population at risk at the beginning of the next interval (j + 1) divided by the population at risk at the beginning of interval j.

$$S_{(j=1,j=4)} = \frac{N'_{04}}{N'_{01}} = \frac{N'_{04}}{N'_{03}} \times \frac{N'_{03}}{N'_{02}} \times \frac{N'_{02}}{N'_{01}}$$

3. The proportion of the original population that remains at risk (i.e., that survives) at the end of a follow-up period with multiple intervals (e.g., at interval 4) is equal to the product of #2 above for each interval.

$$\underbrace{S_{(j=1,j=4)}}_{\text{Cumulative survival}} = \prod_{j=1}^{4} \boxed{4a}\ \frac{N'_{0j+1}}{N'_{0j}} \qquad \underbrace{R_{(j=1,j=4)} = 1 - \left(\prod_{j=1}^{4} \boxed{4b}\ 1 - \frac{I_j}{N'_{0j}} \right)}_{\text{Cumulative risk}}$$

4. Written more generally, this gives us the product limit formula. 4a is the survival in each interval, and 4b is 1 - the incidence proportion for each interval. They are equivalent. The risk in the interval is equal to 1- the grand product of the survival in each interval.

The next video will cover the remaining topics in our outline

- ✓ Recap: cumulative incidence and risk
- ✓ Choosing the appropriate approach to calculating cumulative incidence
  - ✓ Simple cumulative method
  - ✓ Actuarial Method
  - Kaplan-Meier Method
  - Density Method

Berkeley School of Public Health

Calculating cumulative incidence - Part 2

PHW250 F - Jade Benjamin-Chung

Outline

- Recap: cumulative incidence and risk
- Choosing the appropriate approach to calculating cumulative incidence
  - Simple cumulative method
  - Actuarial Method
  - Kaplan-Meier Method
  - Density Method

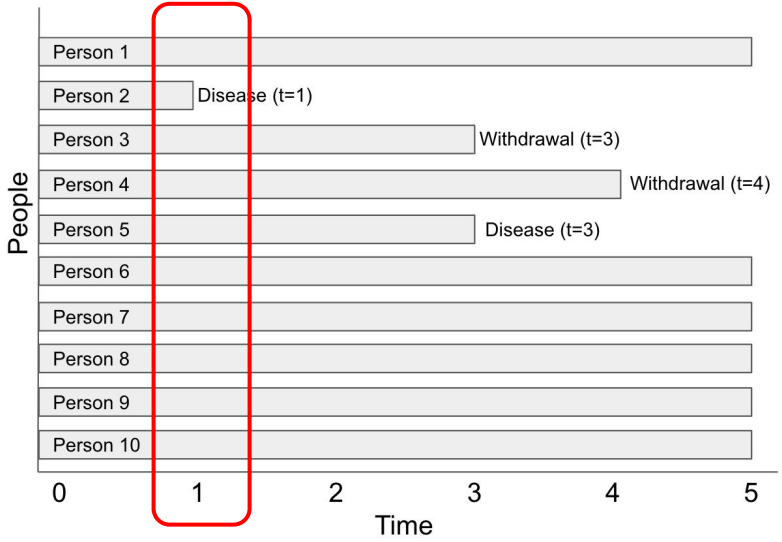
Kaplan-Meier Method

- Appropriate for incomplete follow-up
- Can be used with an open population
- Intuitively, this method involves:
  1. Break the follow-up time into small time intervals
  2. Calculate the risk at the time that each disease event occurs

     The interval risk is a conditional probability — it conditions on whether a person was at risk (alive and not censored) at the event time.
  3. Calculate the **cumulative incidence** that accumulated over all intervals

Kaplan-Meier method

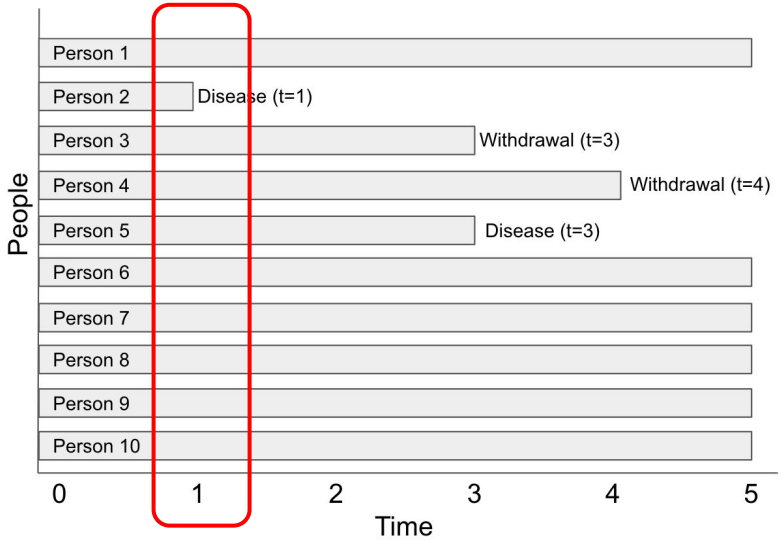
Step 1: Identify the first disease event and calculate the population at risk at the time of the event (N'_j). The person who developed the disease at that time is included in N'_j.



Time j	pop. at risk N'_j	incident disease I_j	interval risk (conditional risk) R_j	interval survival (conditional surv.) S_j
1	10			

Kaplan-Meier method

Step 2: Count the number of incident cases at time **j** (I_j)

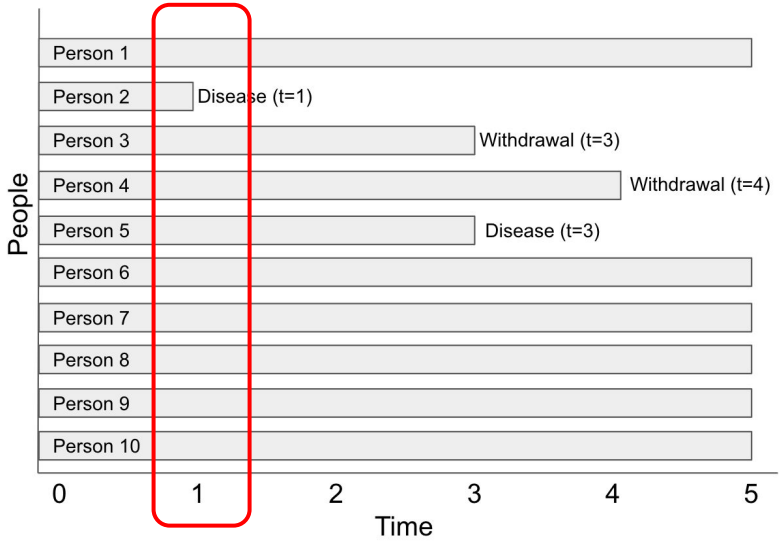


Time j	pop. at risk N'_j	incident disease I_j	interval risk (conditional risk) R_j	interval survival (conditional surv.) S_j
1	10	1		

Kaplan-Meier method

Step 3: Calculate the interval risk ($R_j = I_j / N'_j$)

Keep 3 decimal points!!

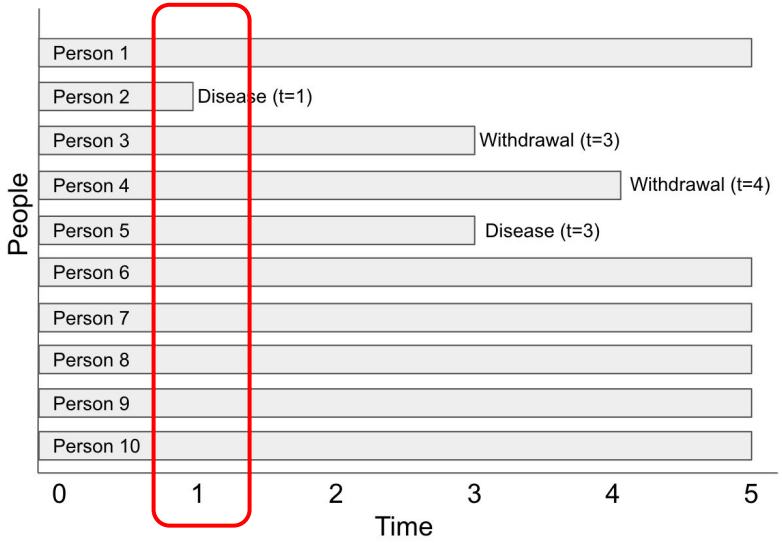


Time j	pop. at risk N'_j	incident disease I_j	interval risk (conditional risk) R_j	interval survival (conditional surv.) S_j
1	10	1	1/10 = 0.100	

Kaplan-Meier method

Step 4: Calculate the interval survival ($S_j = 1 - R_j$)

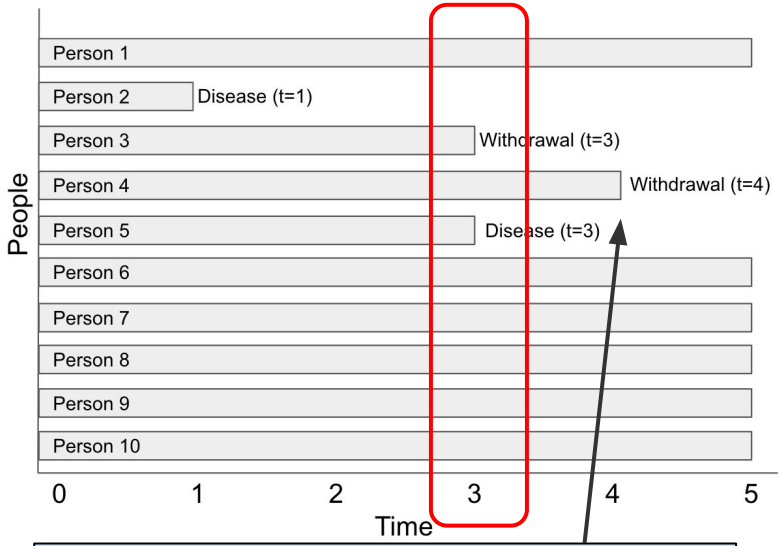
Keep 3 decimal points!!



Time j	pop. at risk N'_j	incident disease I_j	interval risk (conditional risk) R_j	interval survival (conditional surv.) S_j
1	10	1	1/10 = 0.100	0.900

Kaplan-Meier method

Repeat Steps 1-4 for each additional time of disease occurrence



Time j	pop. at risk N'_j	incident disease I_j	interval risk (conditional risk) R_j	interval survival (conditional surv.) S_j
1	10	1	1/10 = 0.100	0.900
3	9	1	1/9 = 0.110	0.889

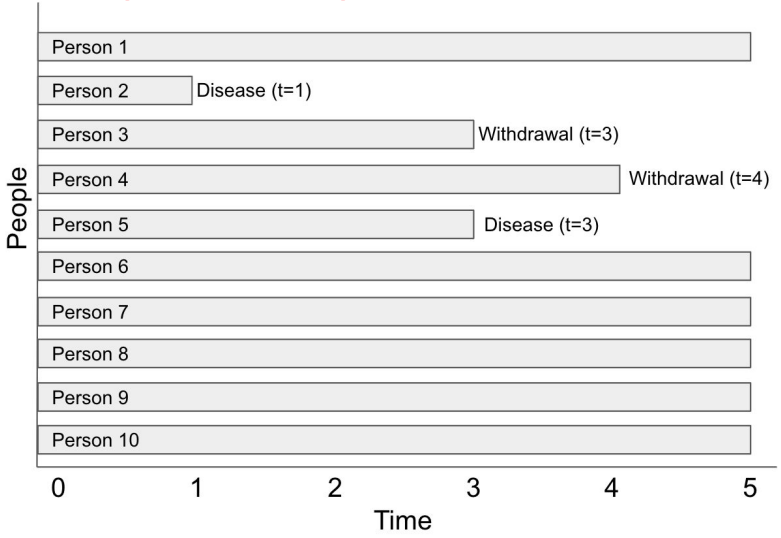
Withdrawals do not count as an "event". This is why we do not include a row for time 4 when a withdrawal occurred but no disease occurred.

Withdrawals are included in **N'_j** up until the time of withdrawal. i.e., the withdrawal at time 3 is included in **N'_3**.

Kaplan-Meier method

Step 6: Calculate the cumulative incidence for the entire follow-up period using the product limit formula: $CI_{(t_0,t_j)} = 1 - \prod_j (1 - R_j) = 1 - \prod_j S_j$

Keep 3 decimal points!! Then round at the end.



Time j	pop. at risk N'_j	incident disease I_j	interval risk (conditional risk) R_j	interval survival (conditional surv.) S_j
1	10	1	1/10 = 0.100	0.900
3	9	1	1/9 = 0.110	0.889

$CI_{(t0,tj)}$ = 1 - (0.900 x 0.889) = 0.1999

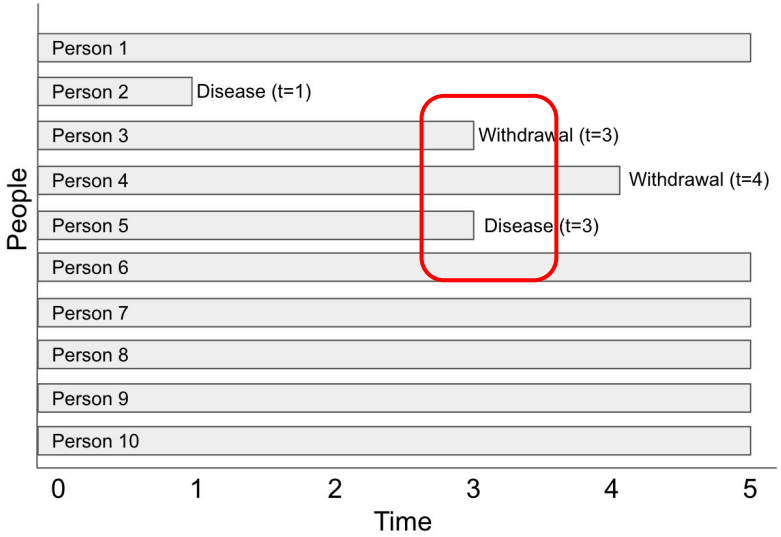
Round to 0.200

Kaplan-Meier method steps

1. Identify the first disease event and calculate the population at risk at the time of the event (**N'_j**). The person who developed the disease at that time is included in **N'_j**.
   - Withdrawals are included in **N'_j** up until the time of withdrawal.
2. Count the number of incident cases at time **j** (**I_j**)
3. Calculate the interval risk (**$R_j = I_j / N'_j$**)
4. Calculate the interval survival (**$S_j = 1 - R_j$**)
5. Calculate the cumulative incidence for the entire follow-up period using the product limit formula: $$CI_{(t_0,t_j)} = 1 - \prod_j (1 - R_j) = 1 - \prod_j S_j$$

Comparing results from the Actuarial and Kaplan-Meier methods

Results from the actuarial and Kaplan-Meier methods will differ when at least one interval has both withdrawal and disease.



Actuarial $R_{t0,tj}$ = 0.206

Kaplan-Meier $R_{t0,tj}$ = 0.200

Comparing results from the Actuarial and Kaplan-Meier methods

Results from the actuarial and Kaplan-Meier methods will differ when at least one interval has both withdrawal and disease.

- The actuarial method assumes that the withdrawal occurred halfway through the interval
- The Kaplan-Meier method does not make this assumption.

Actuarial Method

Interval j	time period t_{j-1}, t_j	pop. at risk N'_{0j}	incident disease I_j	Withdrawals W_j	interval risk (conditional risk) R_j	Interval survival S_j
1	0,1	10	1	0	1 / 10 = 0.100	0.900
2	1,2	9	0	0	0 / 9 = 0.000	1.000
3	2,3	9	1	1	1 / (9 - 1/2) = 0.118	0.882
4	3,4	7	0	1	0 / (7 - 1/2) = 0.000	1.000
5	4,5	6	0	0	0 / 6 = 0.000	1.000

$R_{t0,tj}$ = 1 - (0.900 x 1.000 x 0.882 x 1.000 x 1.000) = 0.206

Kaplan-Meier Method

Interval j	time period t_{j-1}, t_j	pop. at risk N'_{0j}	interval risk (conditional risk) R_j	interval survival (conditional surv.) S_j
1	10	1	1/10 = 0.100	0.900
3	9	1	1/9 = 0.110	0.889

$R_{t0,tj}$ = 1 - (0.900 x 0.889) = 0.200

Outline

- Recap: cumulative incidence and risk
- Choosing the appropriate approach to calculating cumulative incidence
  - Simple cumulative method
  - Actuarial Method
  - Kaplan-Meier Method
  - Density Method

Recap: Risk vs. rates

- **Risk:** the probability that a disease-free individual develops a disease within a specific time period, conditional on that individual not dying from any other disease during the period
  - Has interpretation on the individual level
    - Useful for assessing prognosis of a patient, selecting a treatment strategy, making personal decisions about health behavior
  - **Refers** to a specific period of time
- **Rate:** the average potential for a change in disease status per unit of person-time follow up among disease-free individuals
  - Has no direct interpretation on the individual level
    - Useful for assessing etiologic hypotheses for acute diseases
  - **Includes** the unit of time as part of its definition

Kleinbaum et al., *Epidemiologic Research;* 1982

Density Method

- Used to convert rates to cumulative incidence
- Appropriate for incomplete follow-up
- Can be used with an open population
- Assumes rare disease
- Intuitively, this method involves:
    1. Break the follow-up time into small time intervals
    2. Calculate the incidence density in each interval
    3. Convert the incidence density to an estimate of risk within each interval using a formula that links risks and rates
    4. Calculate the **cumulative incidence** that accumulated over all intervals

Formula to link cumulative incidence and incidence density

$R_j = ID_j \Delta t_j$ = incident cases / person-time at risk x follow-up time
= incident cases / persons at risk

Relationship between risk (R) and incidence density (ID)

Formula to link cumulative incidence and incidence density

$R_j = ID_j \Delta t_j$ = incident cases / person-time at risk x follow-up time
= incident cases / persons at risk

$S_j = 1 - ID_j \Delta t_j$

Relationship between risk (R) and incidence density (ID)

Now calculate S, using S = 1 - R

Formula to link cumulative incidence and incidence density

$R_j = ID_j \Delta t_j$ = incident cases / person-time at risk x follow-up time
= incident cases / persons at risk

$S_j = 1 - ID_j \Delta t_j$

$$CI_{(t_0,t_j)} = 1 - \prod_j (1 - R_j) = 1 - \prod_j S_j$$

Relationship between risk (R) and incidence density (ID)

Now calculate S, using S = 1 - R

As a reminder, here's the product-limit formula

Formula to link cumulative incidence and incidence density

$R_j = ID_j \Delta t_j$ = incident cases / person-time at risk x follow-up time
= incident cases / persons at risk

$S_j = 1 - ID_j \Delta t_j$

$$CI_{(t_0,t_j)} = 1 - \prod_j (1 - R_j) = 1 - \prod_j S_j$$

$$CI_{(t_0,t_j)} = 1 - \prod_j (1 - ID_j \Delta t_j)$$

Relationship between risk (R) and incidence density (ID)

Now calculate S, using S = 1 - R

As a reminder, here's the product-limit formula

Now, let's plug S_j into the product-limit formula

Formula to link cumulative incidence and incidence density

$R_j = ID_j \Delta t_j$ = incident cases / person-time at risk x follow-up time
= incident cases / persons at risk

$S_j = 1 - ID_j \Delta t_j$

$$CI_{(t_0,t_j)} = 1 - \prod_j (1 - R_j) = 1 - \prod_j S_j$$

$$CI_{(t_0,t_j)} = 1 - \prod_j (1 - ID_j \Delta t_j)$$

$$CI_{t_0,t_j} \approx 1 - \exp\left(- \sum_j ID_j \Delta t_j \right)$$

This is the "exponential formula" relating the cumulative incidence and rates.

Relationship between risk (R) and incidence density (ID)

Now calculate S, using S = 1 - R

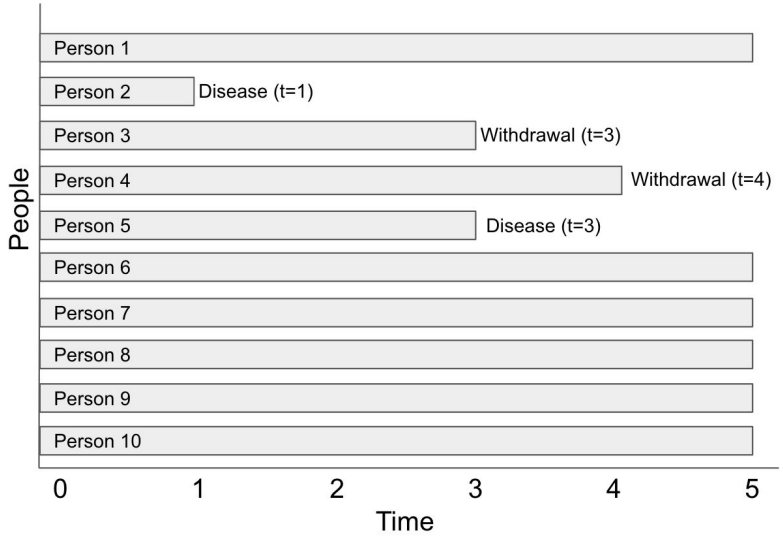
As a reminder, here's the product-limit formula

Now, let's plug S_j into the product-limit formula

Because 1 - x ≈ exp(-x) when x is small, we can re-write the portion of the equation inside the product as shown by substituting $ID_j \Delta t_j$ for x

Density method

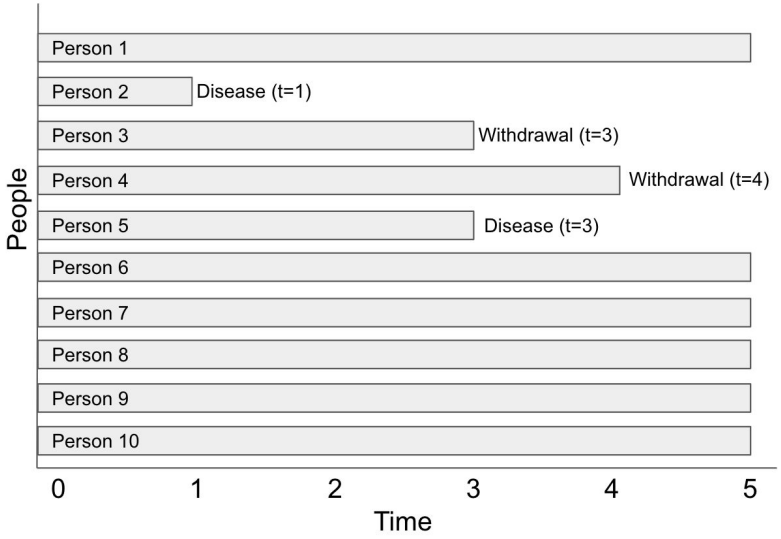
Step 1: Calculate the population at risk at the beginning of each interval (N'_{0j})



Interval j	time t_{j-1}, t_j	Pop. at risk N'_{0j}	incident disease I_j	Interval duration Δt_j	Interval incidence density ID_j	Interval risk $ID_j \Delta t_j$
1	0,1	10				
2	1,2	9				
3	2,3	9				
4	3,4	7				
5	4,5	6				

Density method

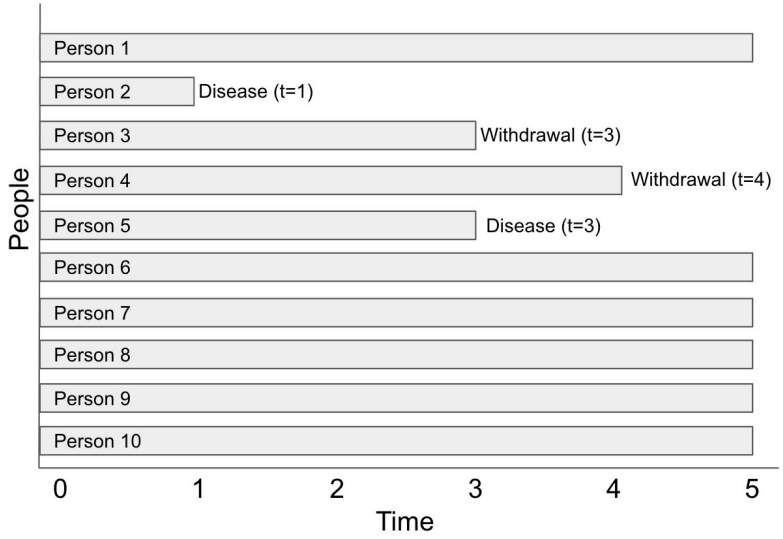
Step 2: Count the number of incident cases in each interval (I_j)



Interval j	time t_{j-1}, t_j	Pop. at risk N'_{0j}	incident disease I_j	Interval duration Δt_j	Interval incidence density ID_j	Interval risk $ID_j \Delta t_j$
1	0,1	10	1			
2	1,2	9	0			
3	2,3	9	1			
4	3,4	7	0			
5	4,5	6	0			

Density method

Step 3: Calculate the duration of follow-up in each interval (Δt_j)

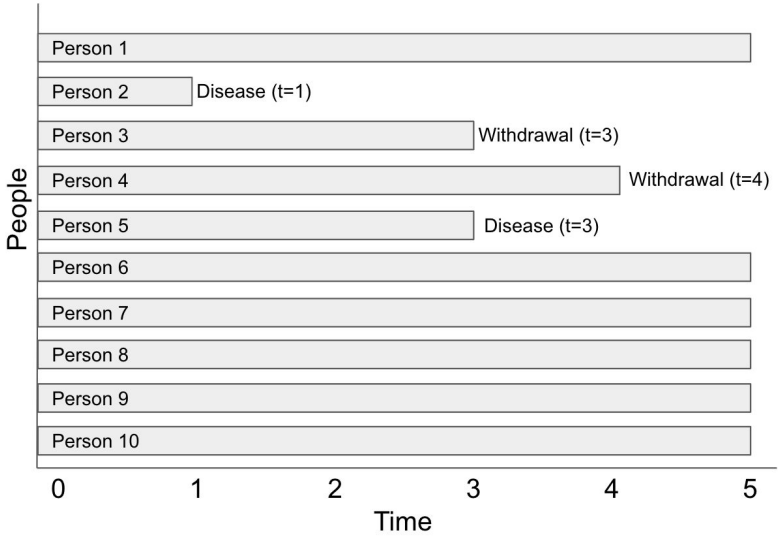


Interval j	time t_{j-1}, t_j	Pop. at risk N'_{0j}	incident disease I_j	Interval duration Δt_j	Interval incidence density ID_j	Interval risk $ID_j \Delta t_j$
1	0,1	10	1	1		
2	1,2	9	0	1		
3	2,3	9	1	1		
4	3,4	7	0	1		
5	4,5	6	0	1		

Density method

Step 4: Calculate the incidence density in each interval: $ID_j = I_j / (N'_{0j}\Delta t_j)$

Keep 3 decimal points!!

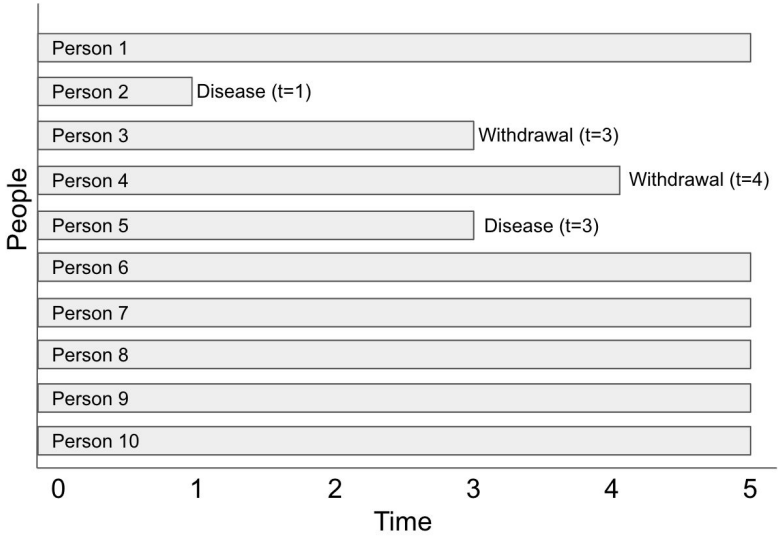


Interval j	time t_{j-1}, t_j	Pop. at risk N'_{0j}	incident disease I_j	Interval duration Δt_j	Interval incidence density ID_j	Interval risk $ID_j\Delta t_j$
1	0,1	10	1	1	1 / 10 = 0.200	
2	1,2	9	0	1	0 / 9 = 0.000	
3	2,3	9	1	1	1 / 9 = 0.111	
4	3,4	7	0	1	1 / 7 = 0.000	
5	4,5	6	0	1	0 / 6 = 0.000	

Density method

Step 5: Calculate the interval risk: $R_j = ID_j \Delta t_j$

Keep 3 decimal points!!



Interval j	time t_{j-1}, t_j	Pop. at risk N'_{0j}	incident disease I_j	Interval duration Δt_j	Interval incidence density ID_j	Interval risk $ID_j \Delta t_j$
1	0,1	10	1	1	1 / 10 = 0.200	0.200 * 1 = 0.200
2	1,2	9	0	1	0 / 9 = 0.000	0.000 * 1 = 0.000
3	2,3	9	1	1	1 / 9 = 0.111	0.111 * 1 = 0.111
4	3,4	7	0	1	1 / 7 = 0.000	0.000 * 1 = 0.000
5	4,5	6	0	1	0 / 6 = 0.000	0.000 * 1 = 0.000

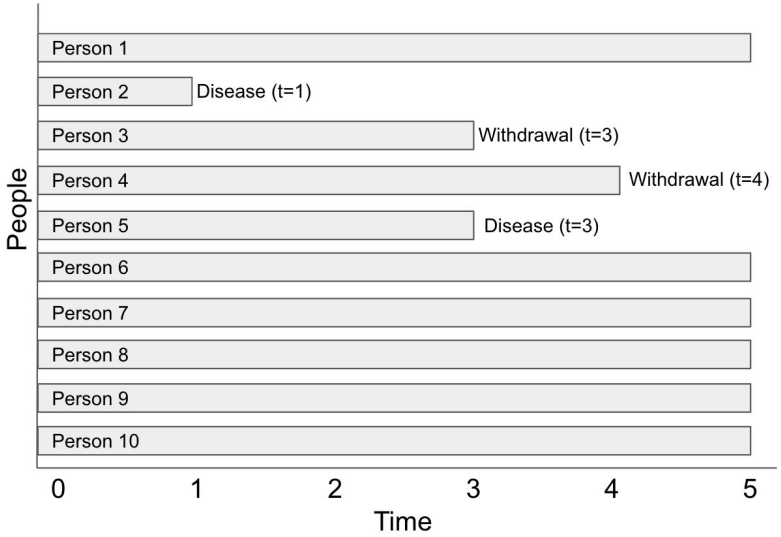
Density method

Step 6: Calculate the cumulative incidence across the entire-follow-up period using the formula:

$$CI_{(t_0,t_j)} \approx 1 - \exp(-\sum_j ID_j \Delta t_j)$$

Note: this formula assumes the interval risk is rare (< 0.1)

Keep 3 decimal points!!



Interval j	time t_{j-1}, t_j	Pop. at risk N'_{0j}	incident disease I_j	Interval duration Δt_j	Interval incidence density ID_j	Interval risk $ID_j \Delta t_j$
1	0,1	10	1	1	1 / 10 = 0.200	0.200 * 1 = 0.200
2	1,2	9	0	1	0 / 9 = 0.000	0.000 * 1 = 0.000
3	2,3	9	1	1	1 / 9 = 0.111	0.111 * 1 = 0.111
4	3,4	7	0	1	1 / 7 = 0.000	0.000 * 1 = 0.000
5	4,5	6	0	1	0 / 6 = 0.000	0.000 * 1 = 0.000

$CI_{(t0,tj)}$ = 1 - exp(-(0.200 + 0.000 + 0.111 + 0.000 + 0.000)) = 0.267

Density method

- Density method estimate = 0.267
- Actuarial method estimate = 0.206
- Kaplan-Meier method estimate = 0.200

What accounts for the difference in the density method estimate?

The disease is not rare (2 out of 5 people developed the disease).

When the disease is not rare, the density method using the exponential formula will not approximate the cumulative incidence.

Density method steps

1. Step 1: Calculate the population at risk at the beginning of each interval (N'_{0j})
2. Step 2: Count the number of incident cases in each interval (I_j)
3. Step 3: Calculate the duration of follow-up in each interval (Δt_j)
4. Step 4: Calculate the incidence density in each interval: $ID_j = I_j / (N'_{0j} \Delta t_j)$
5. Step 5: Calculate the interval risk: $R_j = ID_j \Delta t_j$
6. Step 6: Calculate the cumulative incidence across the entire-follow-up period using the formula:

$$CI_{(t_0,t_j)} \approx 1 - \exp(-\sum_j ID_j \Delta t_j)$$

This formula assumes that the risk in each interval is rare (<0.10).

Summary

- There are three ways to calculate cumulative incidence from individual-level data:
  - **Simple cumulative method:** Appropriate for short time frames and closed populations with no withdrawals
  - **Actuarial method:** Appropriate for open populations with withdrawals when information on disease / withdrawal is available within intervals
  - **Kaplan-Meier method:** Appropriate for open populations with withdrawals when the exact time of disease / withdrawal is available
- If you want to convert incidence density to cumulative incidence, use the **density method**.

Berkeley School of Public Health

Assumptions of cumulative incidence calculations

PHW250B

Assumptions required to calculate cumulative incidence

Method / assumption	**Simple cumulative**	**Actuarial**	**Kaplan-Meier**	**Density**
Uniformity of events and losses within each interval	x	x		x
Independence of censoring and survival	x	x	x	x
No secular trends	x	x	x	x
Assumes population is closed	x	x (for product limit formula)	x (for product limit formula)	x (for product limit formula)
No competing risks	x	x (for product limit formula)	x (for product limit formula)	x (for product limit formula)
Number of events at each event time is a small proportion of the number at risk				x

Assumption 1: Uniformity of events and losses within each interval

- This assumption applies to the Actuarial Method and density method.
- The life table approach assumes that events and losses (withdrawals / deaths) are approximately uniform during each interval.
- If risk changes rapidly within an interval, then estimates that average risk over the interval will not be informative.
- You can adjust the interval size in order to be able to make this assumption.

Assumption 2: Independence of censoring and survival

- This assumption applies to all methods of calculating cumulative incidence.
- Censored individuals have the same probability of the event after censoring as those remaining under observation
  - i.e., censoring is independent of survival
- Example: if patients withdraw from a study because they are sicker than those who do not withdraw, over time, the remaining study population would have patients with decreasing risk of illness, causing incidence to be underestimated.
- This assumption is difficult to make when the disease shares strong risk factors with diseases associated with mortality.

Assumption 3: lack of secular trends

- This assumption applies to all methods of calculating cumulative incidence.
- There are no secular trends in individual characteristics, exposures, or interventions during follow-up that affect the disease.
- Birth cohort and period effects can produce secular trends that bias incidence rate estimates.
- Example: It would not be appropriate to estimate survival from diagnosis of all patients with insulin-dependent diabetes from 1915 through 1935 because this group would include:
  - Patients diagnosed before the introduction of insulin, who had a much lower chance of survival
  - Patients diagnosed after the introduction of insulin, who had a much higher chance of survival
  - It would be more appropriate to calculate incidence rates separately for those time periods

Assumption 4: closed population

- This assumption applies to all methods that use the product-limit formula:
- $R_{(t0, tj)} = 1 - \prod (1 - R_{(tj-1,tj)}) = 1 - \prod (S_{(tj-1,tj)})$

- This is because the incidence proportion is only defined for closed populations.
- In practice, this assumption is often ignored.
- This formula can be used to translate incidence rates from open populations into incidence proportions for closed populations. For example, this is appropriate in a cohort study in which an open population is a subset of a closed population.

Assumption 5: no competing risks

- This assumption applies to all methods that use the product-limit formula:
- $R_{(t0,\, tj)} = 1 - \prod (1 - R_{(tj-1,tj)}) = 1 - \prod (S_{(tj-1,tj)})$
- A competing risk is a risk due to a cause other than the disease under study. For example, in a study of breast cancer risk, a competing risk might be death or withdrawal due to ovarian cancer.
- When there are competing risks, the product limit formula does not hold because:
  - Competing risks may remove additional people between disease onset times, so the population at risk at time t may be smaller than the number of surviving individuals at time t - 1.
  - Population size is not constant in the interval.
- This assumption is made almost universally, but often it is not valid.
- More advanced statistical methods exist for incidence and survival data that better account for competing risks.

Assumption 6: The number of events at each event time is small in proportion to the number at risk at that time

- This assumption applies to the density method because it is required for both the the product-limit formula and the exponential formula:
- $R_{(t0,\, tj)} = 1 - \prod (1 - R_{(tj-1,tj)}) = 1 - \prod (S_{(tj-1,tj)})$
- $R_{(t0,\, tj)} = 1 - e^{(-\sum ID \, * \Delta t)}$
- "Small" means that the incidence rate (I) x $\Delta t < 0.1$
- This assumption is required to substitute in the exponent into the product limit formula. (See Rothman Chapter 3, "Exponential Formula" for full derivation)

Assumptions required to calculate cumulative incidence

Method / assumption	**Simple cumulative**	**Actuarial**	**Kaplan-Meier**	**Density**
Uniformity of events and losses within each interval	x	x		x
Independence of censoring and survival	x	x	x	x
No secular trends	x	x	x	x
Assumes population is closed	x	x (for product limit formula)	x (for product limit formula)	x (for product limit formula)
No competing risks	x	x (for product limit formula)	x (for product limit formula)	x (for product limit formula)
Number of events at each event time is a small proportion of the number at risk				x

Summary of key points

- We covered 4 different methods for calculating cumulative incidence. Each is appropriate for different types of data.
- It is important to assess the assumptions made when calculating cumulative incidence to determine whether they are appropriate in a given study setting. When assumptions are violated, incidence rates will be biased.

Berkeley School of Public Health

Relationships between measures of disease

PHW250B

How are the following related?

- Cumulative incidence and incidence rates
- Prevalence and incidence
- Hazard and incidence

Cumulative incidence and incidence rates

EXHIBIT 2-1 Comparing measures of incidence: cumulative incidence vs incidence rate.

	Cumulative incidence		Incidence rate	
	If follow-up is complete	*If follow-up is incomplete*	*Individual data (cohort)*	*Grouped data (area)*
Numerator	Number of cases	Classic life table Kaplan-Meier	Number of cases	Number of cases
Denominator	Initial population		Person-time	Average population*
Units	Unitless		Time^{-1}	
Range	0 to 1		0 to infinity	
Synonyms	Proportion Probability		Incidence density†	

*Equivalent to person-time when events and losses (or additions) are homogeneously distributed over the time interval of interest.

†In the text, the term *density* is used to refer to the situation in which the exact follow-up time for each individual is available; in real life, however, the terms *rate* and *density* are often used interchangeably.

Szklo 3rd ed.

Cumulative incidence and incidence rates

- We can think of the **incidence rate** as the -(slope/N'_t)
  - slope = - (y2 - y1) / (x2 - x1) = $\Delta N / \Delta t$
  - Incidence rate = $- \Delta N / \Delta t * N'_t$
  - $- \Delta N$ is number of incident cases (negative because the pop. at risk decreases with each incident case)
  - $\Delta t * N'_t$ is person-time
- We can think of **cumulative incidence** as $1 - (N'_t / N'_0) = 1 - e^{(-ID*\Delta t)}$.
  - This is the same formula used in the density method of calculating cumulative incidence.
  - The cumulative incidence is equal to 1 - the point on the solid curved line divided by the y-intercept of that line

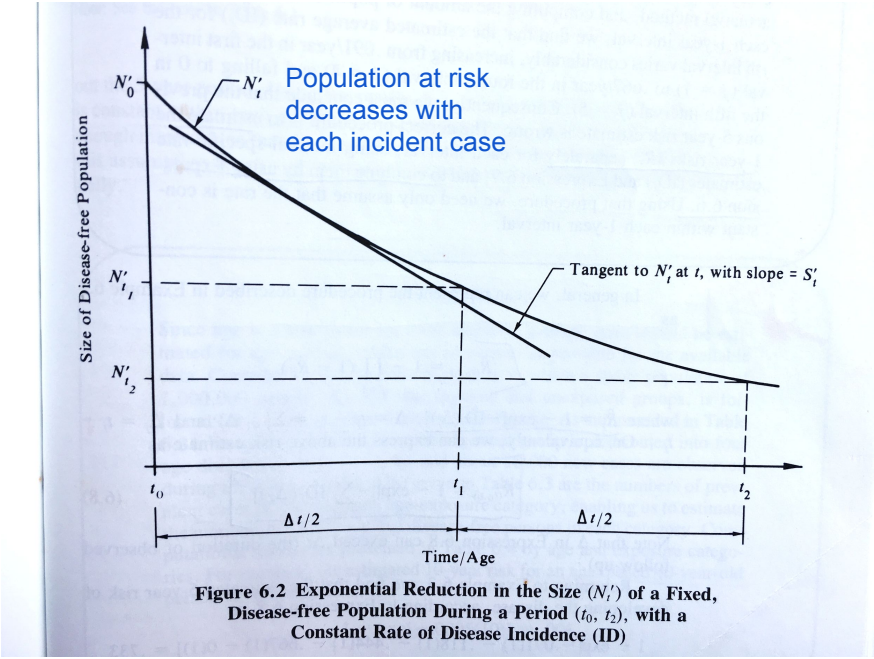


Figure 6.2 Exponential Reduction in the Size (N'_t) of a Fixed, Disease-free Population During a Period (t_0, t_2), with a Constant Rate of Disease Incidence (ID)

Kleinbaum Kupper & Morgenstern. *Epidemiologic Research.* 1982

Prevalence and incidence

- When the population is in a steady state, the **point prevalence** odds approximates the incidence density (ID) times the duration of disease (D)

  $P/(1-P)=ID \times D$

  $P=(IDxD) / (IDxD+1)$

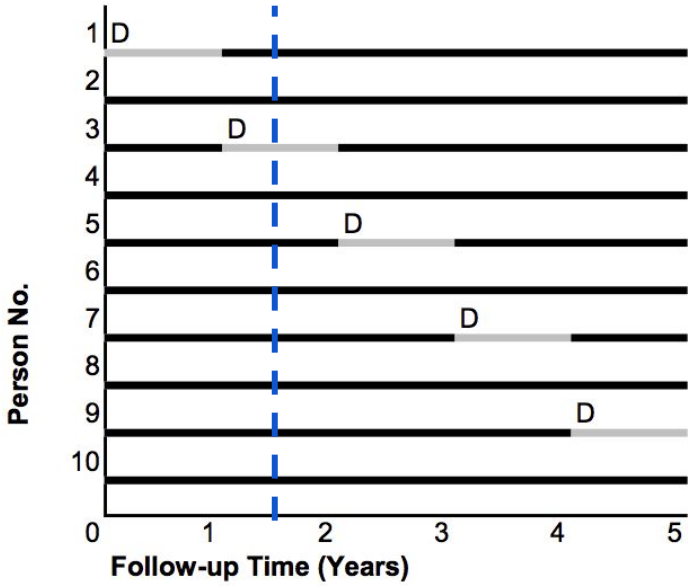
- We can simplify this formula if the disease is rare ($P<0.1$):
  - $P \sim ID \times D$
  - This equation is an approximate because when $P<0.1$, $P/(1-P) \sim P$

Prevalence and incidence formula - why does the population need to be in a steady state?

- Because in a steady state, incident cases (inflow) = recovered cases (outflow)
- Inflow = incidence rate (I) * susceptible population (N_S)
- Outflow = recovery rate (r) * diseased, not at-risk pop (N_D)
- Under steady state,
  - $I \times N_s = r * N_D$
  - Prevalence odds = $N_D/N_S = I * (1/r)$
  - and 1/r = duration
  - $P/(1-P) = ID \times D$

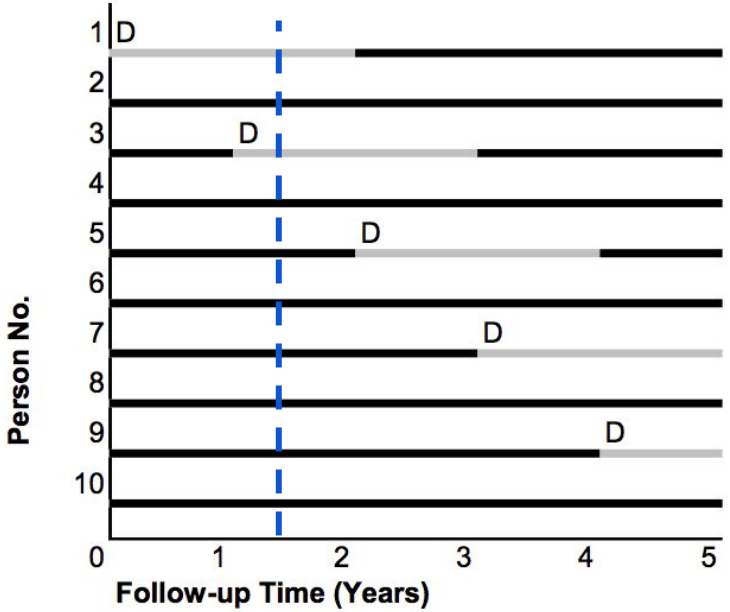
Relationship between incidence and prevalence depends on duration (assume recovery after disease)

Scenario 1: duration = 1 year



Prevalence from Year 1-2: 1/10 =0.1
Incidence density from Year 1-2: 1/10 = 0.1 person-yrs
Under rare disease assumption:
P = ID x D = 0.1 x 1 = 0.1

Scenario 2: duration = 2 years



Prevalence from Year 1-2: 2/10 =0.2
Incidence density from Year 1-2: 1/9 = 0.11 person-yrs
Under rare disease assumption:
P = ID x D = 0.11 x 2 = 0.22 ~ 0.2

Hazard and incidence

- **Hazard:** The instantaneous potential for change in disease status per unit of time at time t relative to the size of the candidate (i.e., disease-free) population at time t
- It is also called “instantaneous conditional incidence” or the “force of morbidity (or mortality)”
- How is hazard different from incidence?
  - Hazard is an instantaneous rate
  - Incidence density is an average rate over a period of time
- Hazard cannot be directly calculated because it is defined for an infinitely small time interval
- Statistical models can be used to estimate the hazard function over time
- It is frequently used in cohort studies where the outcome is survival time or time to event

Summary of relationships between measures of disease

- **Cumulative incidence and incidence density:**
  - CI = 1 - e(-ID*Δt).
- **Prevalence and incidence density**
  - Under steady state
    - P/(1-P)=ID x D
    - P=(IDxD) / (IDxD+1)
  - Under steady state + disease is rare
    - P = ID x D
- Hazard is the instantaneous potential for change in disease status per unit of time at time t. It cannot be directly calculated because it is defined for an infinitely small time interval.

Berkeley School of Public Health

Indirect standardization

PHW250 F - Jade Benjamin-Chung

Standardization

- **Crude rates** in different populations might not be comparable due to differences in those populations (e.g., different age distributions)
- **Standardized or adjusted rates:** have been transformed statistically to allow for direct comparison between populations
  - Standardization allows us to adjust for one characteristic at a time (focus of this video)
  - Multivariate models allow you to adjust for multiple characteristics at a time (not covered in detail in this class)

Recap: Direct Standardization

Direct standardization:

1. Obtain population counts stratified by a covariate for an outside reference/standard population
2. Apply the stratified rates from your study populations to the stratified population counts in the reference population
3. Calculate the **expected number of people with the disease** in each study population had they had the stratified population counts of the reference population (a counterfactual concept)
4. **Calculate adjusted rate:** total expected outcomes / total reference population
5. **Calculate adjusted RR or RD** using adjusted rates for the exposed and unexposed.

   Note: The value of the adjusted rate in Step 4 will depend on the reference population chosen.

Direct vs. Indirect Standardization

Direct standardization:

1. Obtain population counts stratified by a covariate for an outside reference/standard population
2. Apply the stratified rates from your study populations to the stratified population counts in the reference population
3. Calculate the **expected number of people with the disease** in each study population had they had the stratified population counts of the reference population (a counterfactual concept)
4. **Calculate adjusted rate:** total expected outcomes / total reference population
5. **Calculate adjusted RR or RD** using adjusted rates for the exposed and unexposed.

   Note: The value of the adjusted rate in Step 4 will depend on the reference population chosen.

Indirect standardization:

1. Obtain death or disease rates stratified by a covariate for an outside reference/standard population
2. Apply the stratified rates from the reference population to the stratified population counts in your study populations
3. Calculate the **expected number of people with the disease** in each study population had they had the stratified rates of the reference population (a counterfactual concept)
4. **Calculate total observed outcomes** in the study population and **total expected outcomes** applying the reference population's rates to the study population counts
5. **Calculate standardized incidence/ prevalence /mortality ratio (SIR / SPR / SMR):** observed number diseased / expected number diseased

Standardized incidence/prevalence/mortality ratio

- SIR/SPR/SMR = $\frac{\text{total observed outcomes}}{\text{total expected outcomes}}$
- Ratio of the observed number of outcomes in the study population to the expected number of outcomes if the study population had the same stratified rates as the reference population
- **Answers question:** how do rates in the study population compare to rates in the reference population?
  - SMR< 1: the rate is lower in the study population than the reference population
  - SMR = 1: the rates are the same
  - SMR> 1: the rate is higher in the study population than in the reference population

Step 1: Obtain rates from standard population

	Study group A				Study group B				Standard population rates
Age group	**N**	**Deaths**	**Rate**	**Expected deaths**	**N**	**Deaths**	**Rate**	**Expected deaths**	
< 40 years	100	10	10%		500	50	10%		12%
≥ 40 years	500	100	20%		100	20	20%		50%
Total	**600**	**110**	18.3%		**600**	**70**	11.7%		
			SMR:				SMR:		

Szklo & Nieto 3rd Ed., Table 7.8

Step 2: Apply standard pop. rates to study populations

	Study group A				Study group B				Standard population rates
Age group	**N**	**Deaths**	**Rate**	**Expected deaths**	**N**	**Deaths**	**Rate**	**Expected deaths**	
< 40 years	100	10	10%	100 * 12%	500	50	10%	500*12%	12%
≥ 40 years	500	100	20%	500 * 50%	100	20	20%	100*50%	50%
Total	**600**	**110**	18.3%		**600**	**70**	11.7%		
			SMR:				SMR:		

Szklo & Nieto 3rd Ed., Table 7.8

Step 3: Calculate expected # of deaths

	Study group A				Study group B				Standard population rates
Age group	**N**	**Deaths**	**Rate**	**Expected deaths**	**N**	**Deaths**	**Rate**	**Expected deaths**	
< 40 years	100	10	10%	100 * 12% = 12	500	50	10%	500*12%=60	12%
≥ 40 years	500	100	20%	500 * 50% = 250	100	20	20%	100*50%=50	50%
Total	**600**	**110**	18.3%	12+250 = **262**	**600**	**70**	11.7%	60+50=**110**	
			SMR:				SMR:		

Szklo & Nieto 3rd Ed., Table 7.8

Step 4: Calculate SMRs

	Study group A				Study group B				Standard population rates
Age group	**N**	**Deaths**	**Rate**	**Expected deaths**	**N**	**Deaths**	**Rate**	**Expected deaths**	
< 40 years	100	10	10%	100 * 12% = 12	500	50	10%	500*12%=60	12%
≥ 40 years	500	100	20%	500 * 50% = 250	100	20	20%	100*50%=50	50%
Total	**600**	**110**	18.3%	12+250 = **262**	**600**	**70**	11.7%	60+50=**110**	
			SMR:	110/262 = 0.42			SMR:	70/110=0.64	

Interpretation of SMR for group A: the mortality rate in Study group A is lower than in the standard population

Interpretation of SMR for group B: the mortality rate in Study group B is lower than in the standard population

Szklo & Nieto 3rd Ed., Table 7.8

Standardized illness/mortality ratio

- When comparing more than one study population to the standard population rates, the SMRs are using different standards (since the study populations are serving as the standard).
- Thus, it is usually not appropriate to compare SMRs or SIRs between study groups.

Szklo & Nieto 3rd Ed.

Can we compare SMRs for group A and B?

	Study group A				Study group B				Standard population rates
Age group	**N**	**Deaths**	**Rate**	**Expected deaths**	**N**	**Deaths**	**Rate**	**Expected deaths**	
< 40 years	100	10	10%	100 * 12% = 12	500	50	10%	500*12%=60	12%
≥ 40 years	500	100	20%	500 * 50% = 250	100	20	20%	100*50%=50	50%
Total	**600**	**110**	18.3%	12+250 = **262**	**600**	**70**	11.7%	60+50=**110**	
			SMR:	110/262 = 0.42			SMR:	70/110=0.64	

- In this example, two study groups have identical age-specific rates, but their age distributions are different.
- Applying the reference population rates to these study groups will yield expected counts that are based on different weights. Thus, they cannot be directly compared.

Szklo & Nieto 3rd Ed., Table 7.8

Summary of key points

Direct standardization:

- Use stratified rates from your study population
- Use stratified population counts from a standard population
- Calculate standardized rate and adjusted measure of association
- Provide information about the burden of disease

Both:

- Invoke the concept of counterfactuals
- Are used to adjust for confounding by a single variable that is associated with the outcome (disease/death)

Indirect standardization:

- Use stratified population counts from your study population
- Use stratified rates from a standard population
- Calculate SMR (or SIR or SPR)
- Useful when stratified rates are not available for your study population or when the number of observations in each stratum is too small to estimate stable stratified rates
- Do not provide information about the burden of disease
- Most popular in occupational epidemiology