PHW250B Week 3 Reader

Topic 1: Age, Period, and Cohort Effects

Topic 2: Calculating Incidence

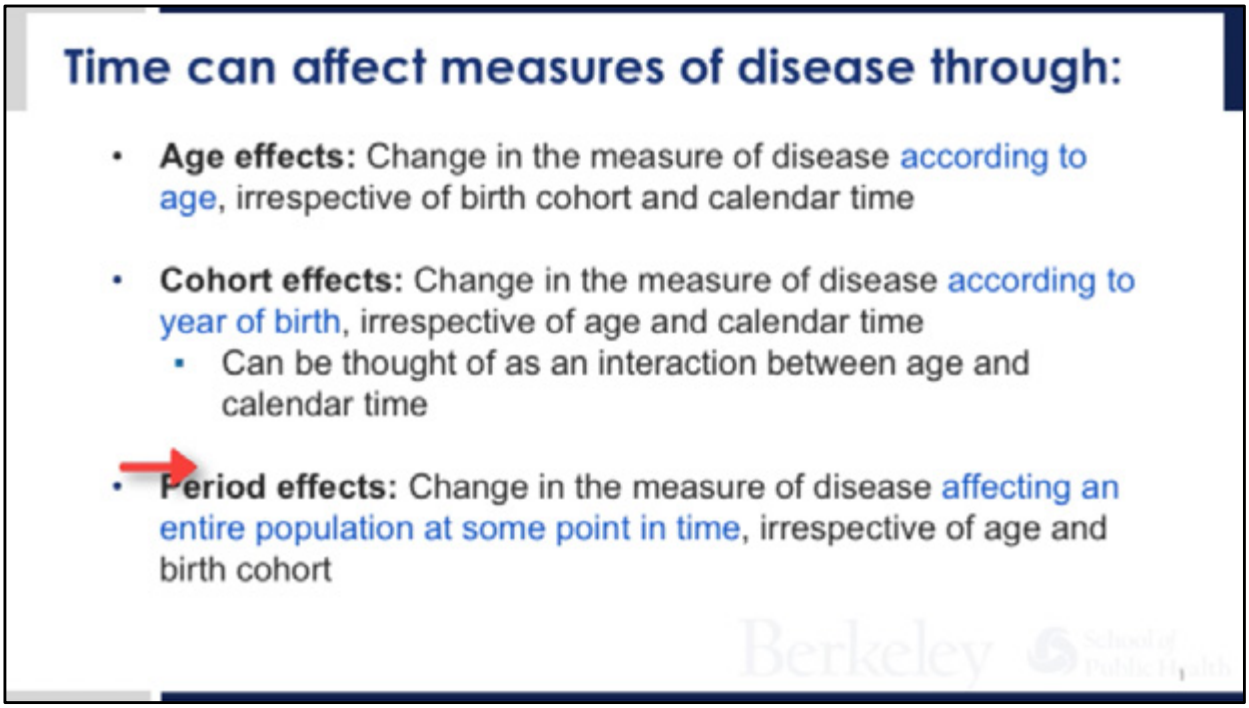
Topic 3: Relationships Between Measures of Disease

Topic 4: Indirect Standardization

Journal Club

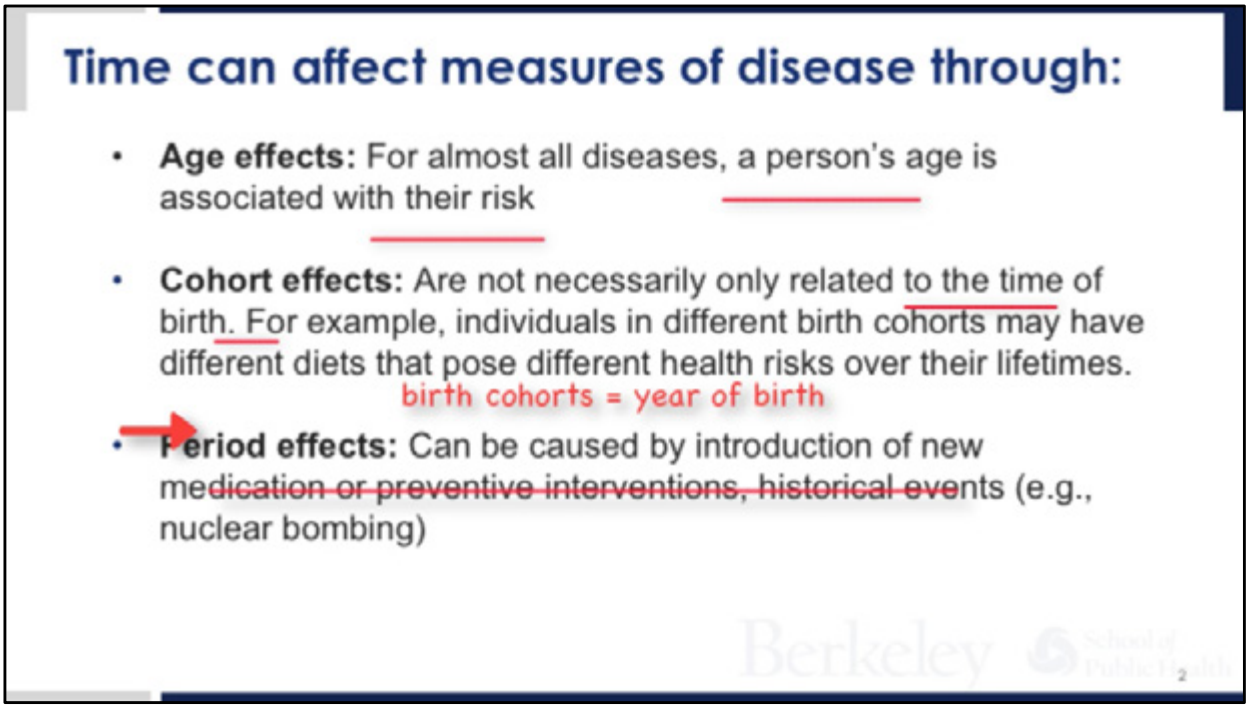
Lecture 3.1.1: Age, Period, and Cohort Effects

In this lesson, we're going to discuss age, period, and cohort effects, three related and important concepts that apply to the study of populations and are often very important for epidemiologists to understand when trying to disentangle causes of disease.



Let's define each of them, but I think they'll make the most sense to you when you see pictures and figures describing real examples. But let me first just put down in words here the definitions of these three different effects, age, cohort, and period effects. So age effects are defined as the change in the measure of disease according to age, irrespective of birth, cohort, and calendar time.

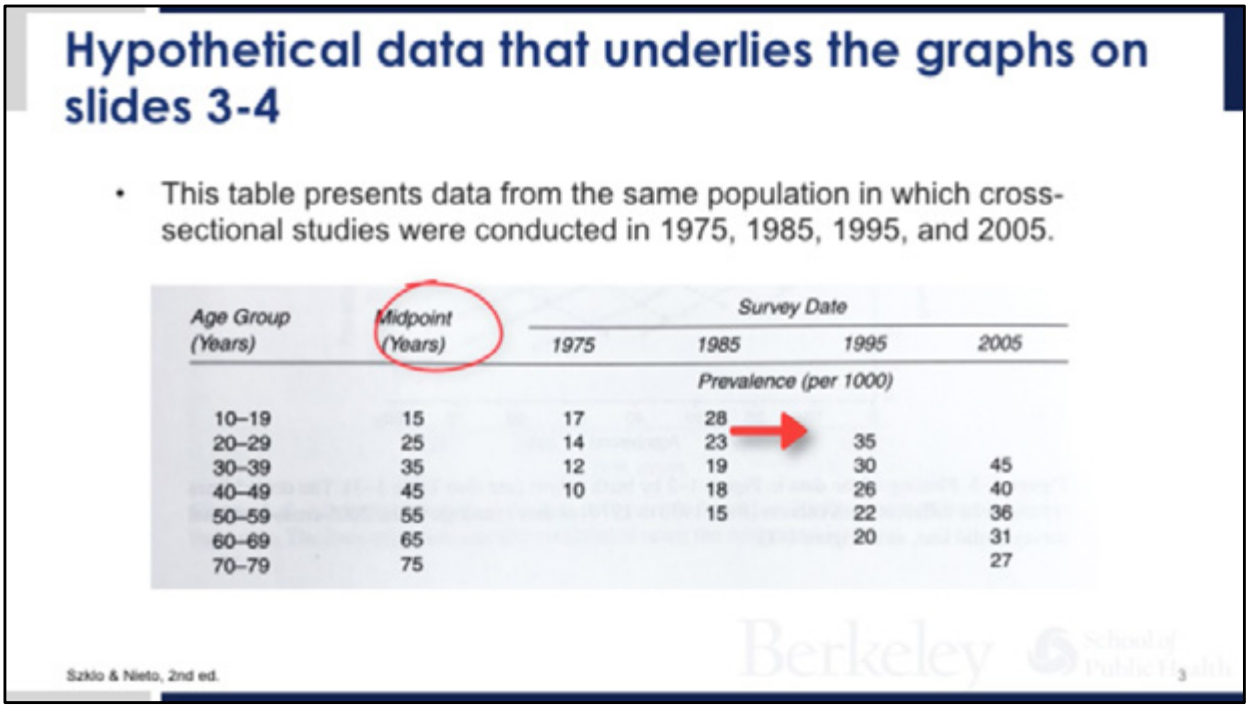
Cohort effects are the change in the measure of disease that occur according to the year of birth, irrespective of age and calendar time. These can be thought of as an interaction between age and calendar time. And finally, period effects are the change in the measure of disease affecting an entire population at some point in time, irrespective of age and birth cohort. I think these will make more sense as we do examples, first a hypothetical example and then some actual examples.



For age effects, for almost all conditions, a person's age is associated with risk. So this doesn't necessarily mean that risk goes up with age. For some conditions, risk will go down with age. But if there's a relationship between age and the rate or prevalence of disease, we refer to that as an age effect.

Secondly, cohort effects are not necessarily only related to the time of birth. And by time of birth, I mean year of birth-- 1960, 1970, 1980, and so forth. For example, individuals in different birth cohorts-- and by birth cohorts, I mean the year or the year group in which person a is born-- they might have had different diets that occurred over time that affected those groups differently and therefore pose different health risks over their lifetimes. For example, someone born in 1950 and living for 80 years experiences a different diet than someone born in 1990 and lives for 50 years. We'll see examples of this as we go forward.

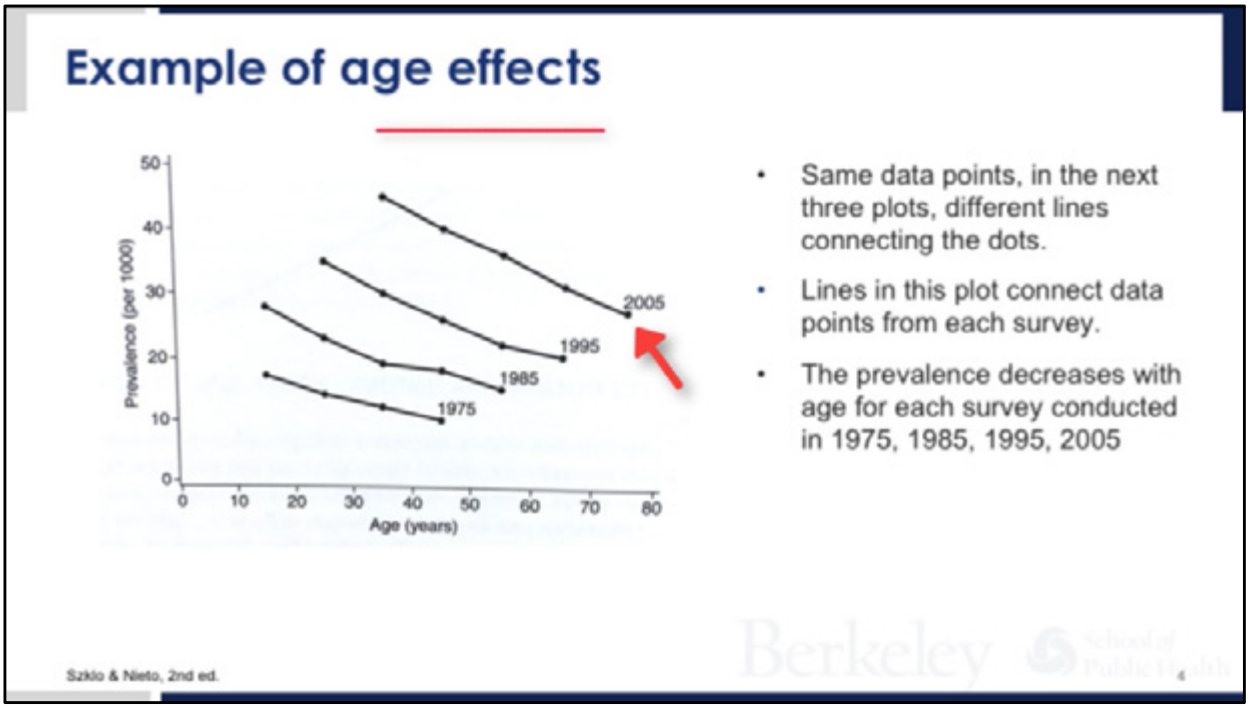
Period effects can be caused by the introduction of new medications or preventive interventions or historical events, even, for example, huge things like nuclear bombs. These are big population events that cause period effects in our observation of the rates of disease.



This is a table of hypothetical data describing multiple surveys done on a population in different years. Imagine that in 1975, 1985, 1995, and 2005, a survey was done of a population. That might be the state of California, for example. And in each of those years, there were measurements made on all the ages of people in the population.

In the first row, the people who were between the ages of 10 and 19 are studied, and they're referred to by the midpoint of their age group. So we'll refer to that row as the age 15 row. And in 1975, when the survey was done of the population, there was a 17% prevalence of disease. In 1985, there was a 28% prevalence of disease.

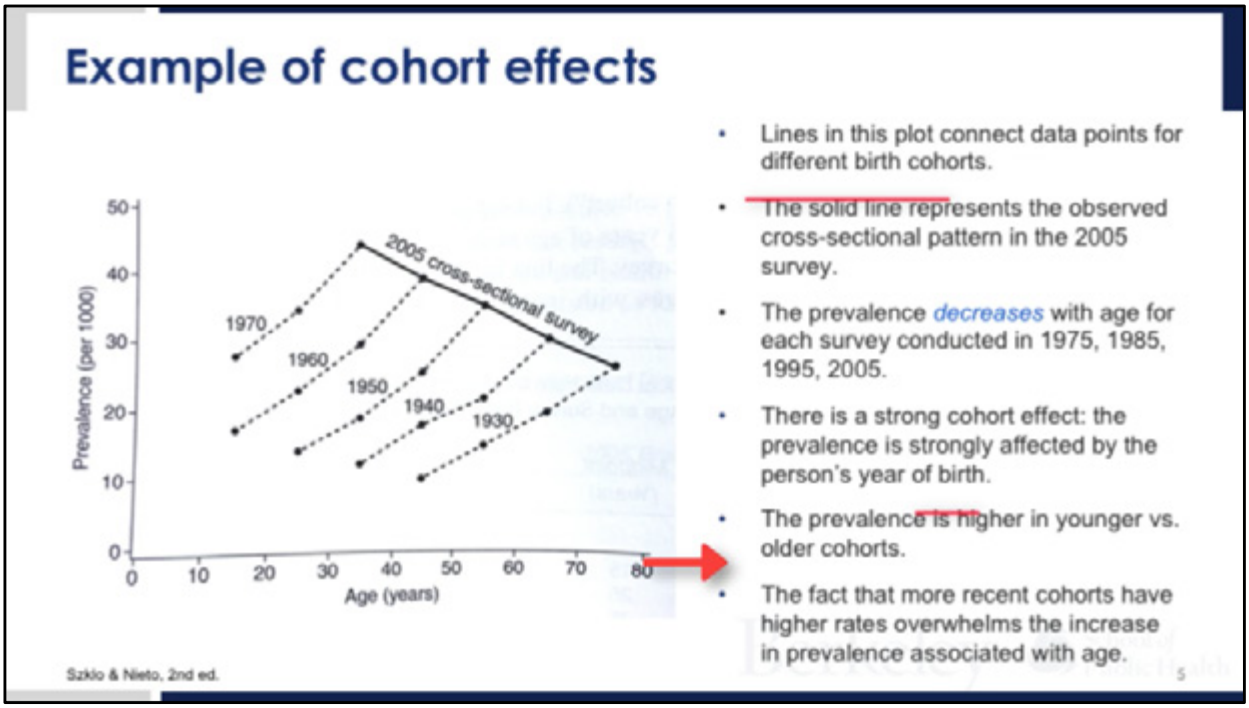
For the people aged 20 to 29, we'll refer to them by their midpoint, so that would be 25 years of age. And in 1975, that age group of the population had 14% prevalence of disease. And in 1985, it had 23% prevalence of disease. And in 1995, it had 35% prevalence of disease. And so forth. We're now going to plot the data in various ways to see what we can understand about age, period, and cohort effects.



First, let's look at age effects. If we take the data from the 1975 survey and we plot it, we see that the prevalence of disease in the people who were age 15-- the mid-point age, 15-- was about 18%. And in the mid-point age 25 group, the prevalence was about 15%. And in the mid-point age 35, the prevalence was about 12%. And in the 45-year-old mid-point age group, the prevalence was about 10%.

That's the description point by point for the survey done in 1975 for all the ages in the population. That's then repeated in each of the surveys. All the ages are measured in each of the years in which the survey is done. We can see that what's happening to the prevalence of disease within each survey year, the prevalence of disease goes down with age.

For example, the 2005 survey, the prevalence of disease is highest in the youngest group that's available for measure. That's the 35-year-olds. And it's lowest in the oldest group, the 75-year-olds. So hopefully this age affect is very apparent in this graph.



Now, let's do something rather clever with the data. Let's remove the lines that we had on the prior graph but now plot the same points again, but we're going to connect them in a different way. Rather than the connections we made within a survey year, we're now going to draw different lines and connect the ages within different birth cohorts.

Everybody born in 1970 tracked through their various ages using these surveys has their data plotted. People born in 1970, when they were age 15, their prevalence was about 28%. When they were age 25, their prevalence was about 35%. And so on and so forth across the different birth cohort experiences across the ages.

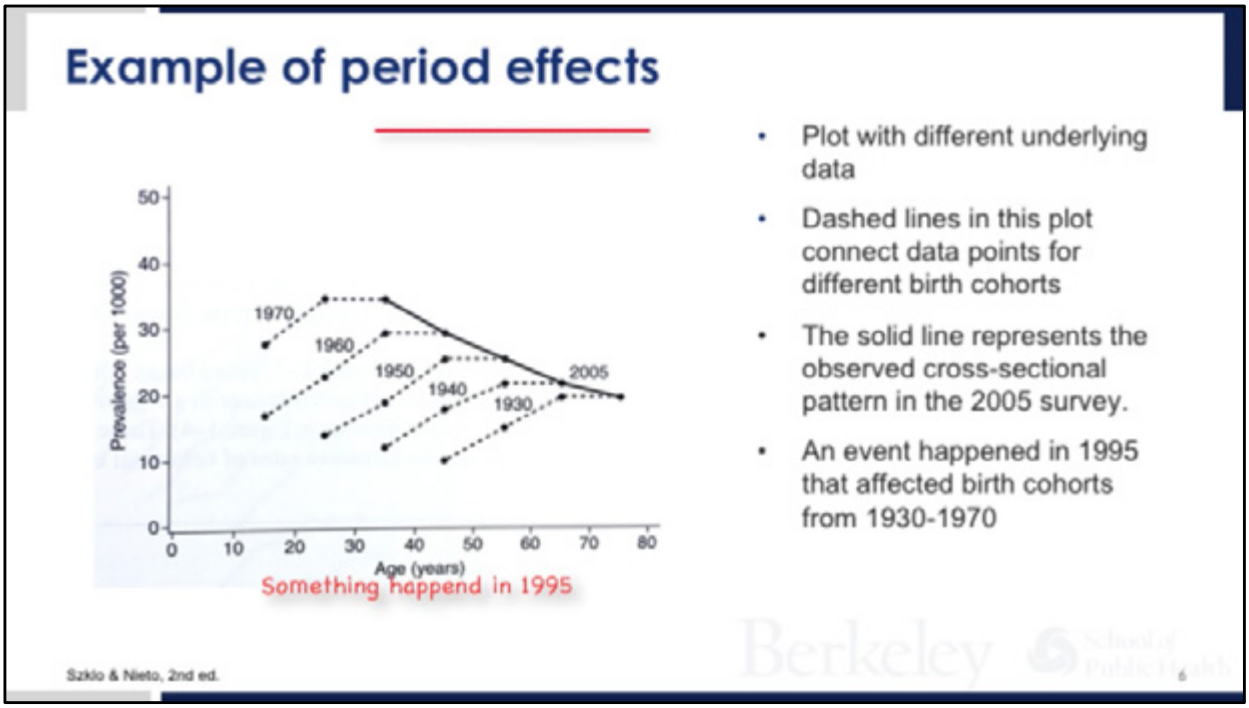
We've taken the data from the previous slide, which was the surveys done in each year. But now we're connecting the same group of people over time. Still showing on this graph was the original cross-sectional survey in the solid line for 2005. So that comes from the prior graph, just to remind you how these are all connected.

We're connecting the data points for different birth cohorts. As I said, the solid line represents the observed cross-sectional pattern. What we can see here is that the prevalence decreases with age for each survey conducted in '75, '85, '95, and 2005.

That was the prior slide.

But now there's a strong cohort effect. The prevalence is strongly affected by the person's year of birth. So look at people born in 1930 versus 1940 versus 1950 versus 1960 versus 1970. Now what we see is that as they get older within each birth cohort, the prevalence of disease increases. Here, what we see is that the prevalence is higher in younger versus older cohorts.

So this is a different conclusion than we had on the prior slide, when it looked like the prevalence was decreasing with age. The fact that more recent cohorts have higher rates overwhelms the increase in prevalence that's associated with age. Hopefully, you can see by looking at these two graphs together how age and cohort effects can be intertwined with each other.



Now, let's move on and talk about period effects. In this graph, we're using different underlying data. You're going to get confused if you think we're re-plotting the data from the prior slides. We're not doing that. We're using related data but with a twist that you'll see here related to the period effects.

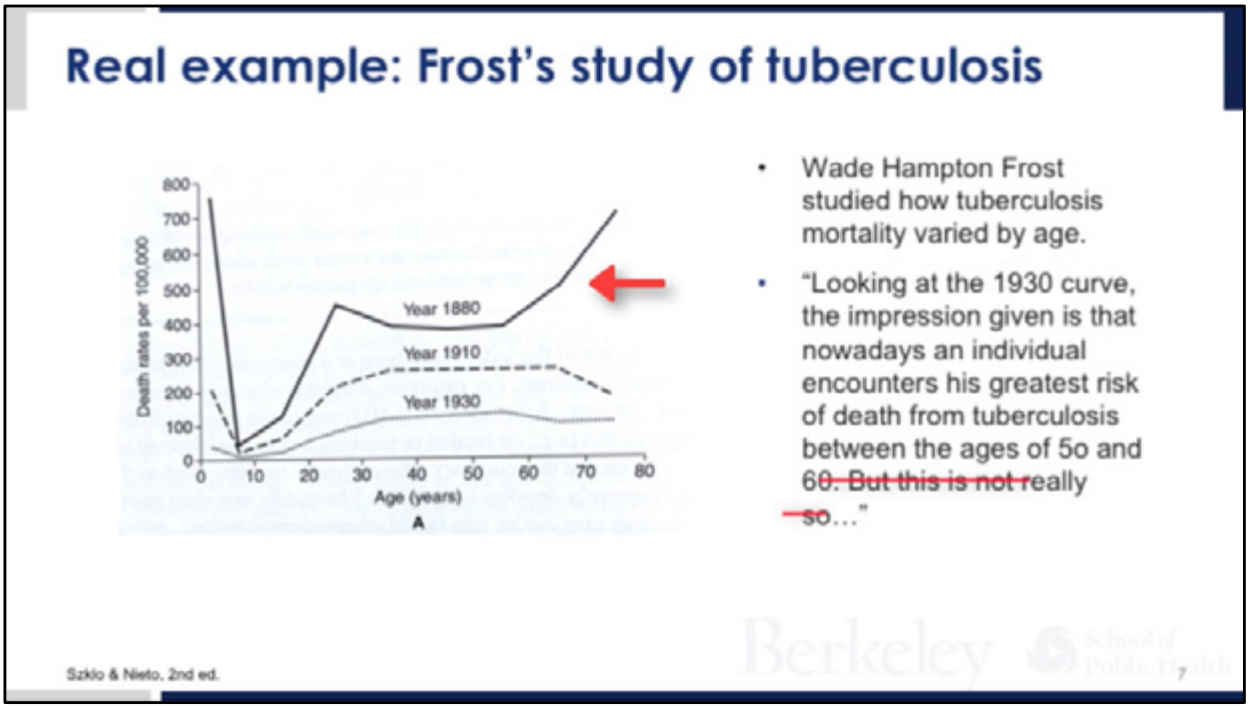
In this graph, the dash lines are again connecting data points for different birth cohorts. The solid line represents the observed cross-sectional pattern in the 2005 survey. And what happened here was that something occurred in 1995 that affected the birth cohorts from 1930 to 1970. So let's disentangle that and see what I'm saying here.

The people born in 1970 at age 10 had a prevalence of disease about 25%. Then in age just above 20, they had a prevalence of about 33%. But notice what happens to the 1970 birth cohort. At about age 25, their curve levels off. That should be obvious to you that it's not increasing after the prevalence rose to its highest level. At 1970, about 25 years later, something happened that leveled off the age effect and what was changing in that cohort.

Now look at the 1960 group. At about age 35, there was a leveling. That would, again,

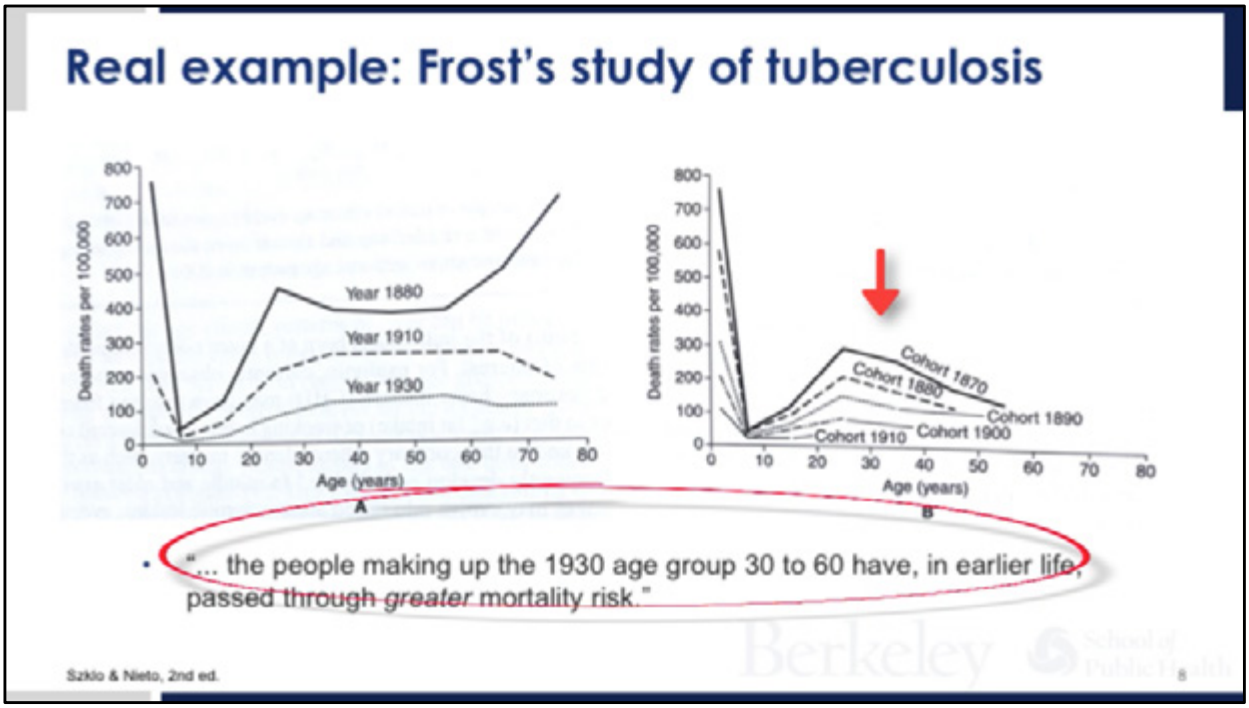
be the year 1995. Now look at the 1950 group. When they were 45, something happened. And what year was that?

Well, when the 1950 birth cohort is age 45, that's the year 1995. Similarly for the 1940 and 1930 groups, something happened in 1995 that affected the prevalence of disease in each cohort. And this is called a period effect. Something about the period, 1995, changed the disease experience of the population.



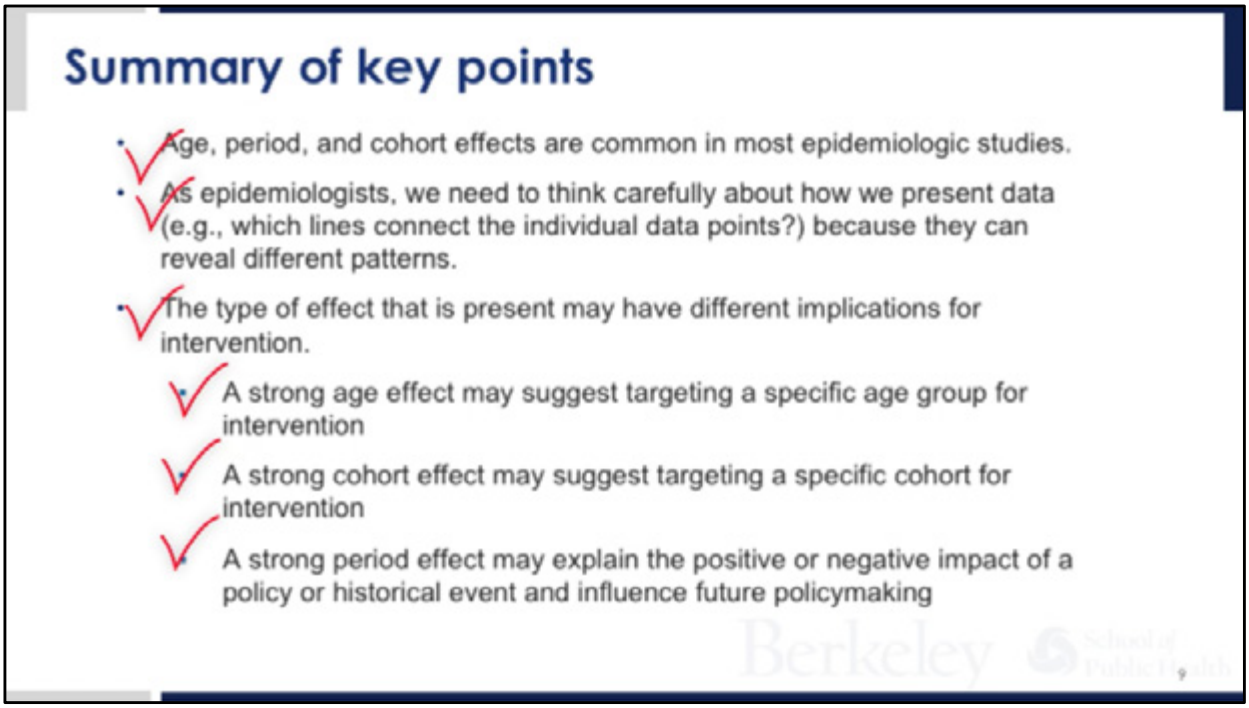
Now let's work with some actual data from real studies. This is a clever example from the work of Wade Hampton Frost about tuberculosis mortality. And what we see here is an impression, just taking from his words, "Looking at the 1930 curve," it looked like "an individual encounters his greatest risk of death from tuberculosis between the ages of 50 and 60. But this is not really so."

And the reasons it's not really so is because of these concepts we've studied today about age and period effects. What's happening in this graph is, the surveys done in 1930, 1910, and 1880 plotted the different rates of tuberculosis mortality for different ages in those different years of the surveys. But in actuality, what was really happening was plotting the data a different way.



The slide on the left is what we just saw, the surveys done in the different years. But in fact, if we look at the plots of the experience of each cohort-- in other words, the cohorts from 1900, 1910, 1890, 1880, and 1870-- we see something very different. The people making up the 1930 age group, 30 to 60, had in their earlier life passed through greater mortality risk.

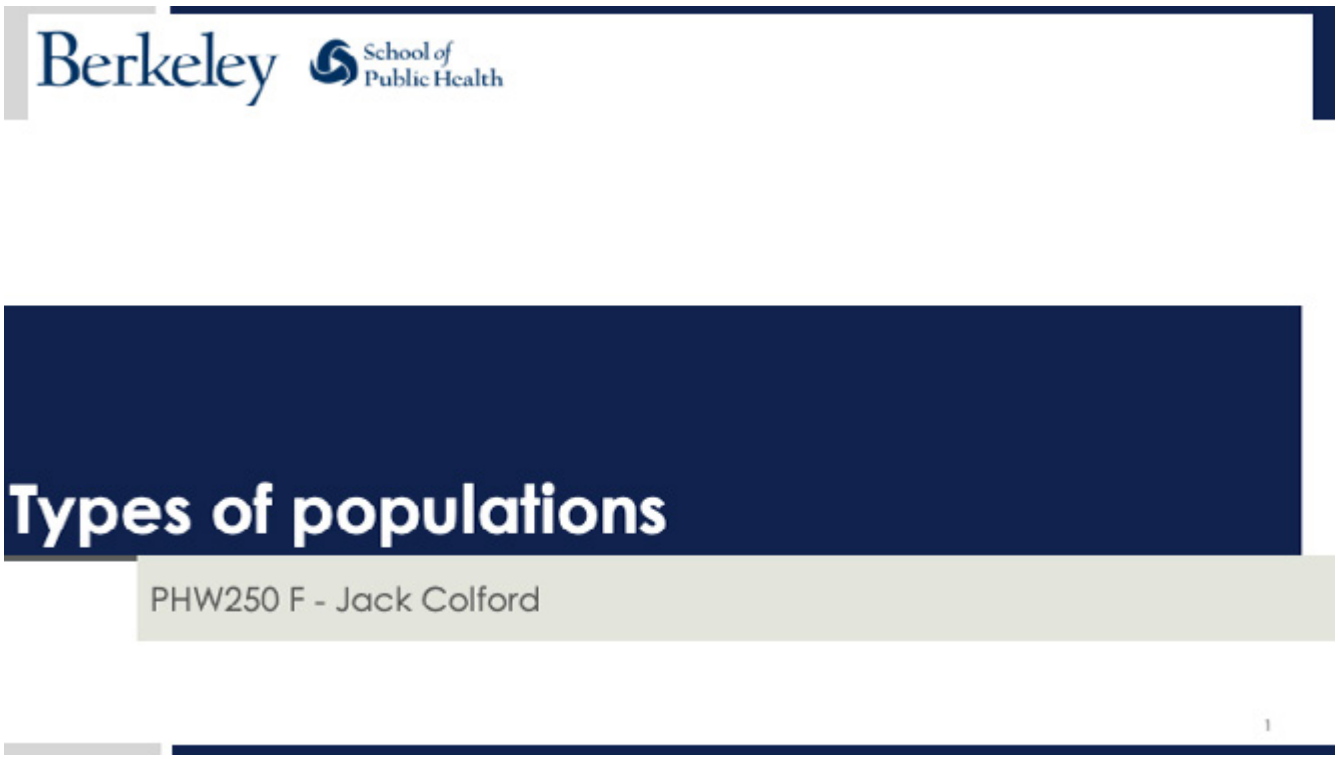
So now we have a very different interpretation, because we see that the experience of tuberculosis mortality was different for the different cohorts. There's still an age effect. But what we're seeing really strongly here is a cohort effect. So hopefully the juxtaposition of these plots of actual data from the same study help you to see how you can be misled by an examination that doesn't take into account all of the various things going on in the population-- in this case, the changes in the different cohorts that were occurring.



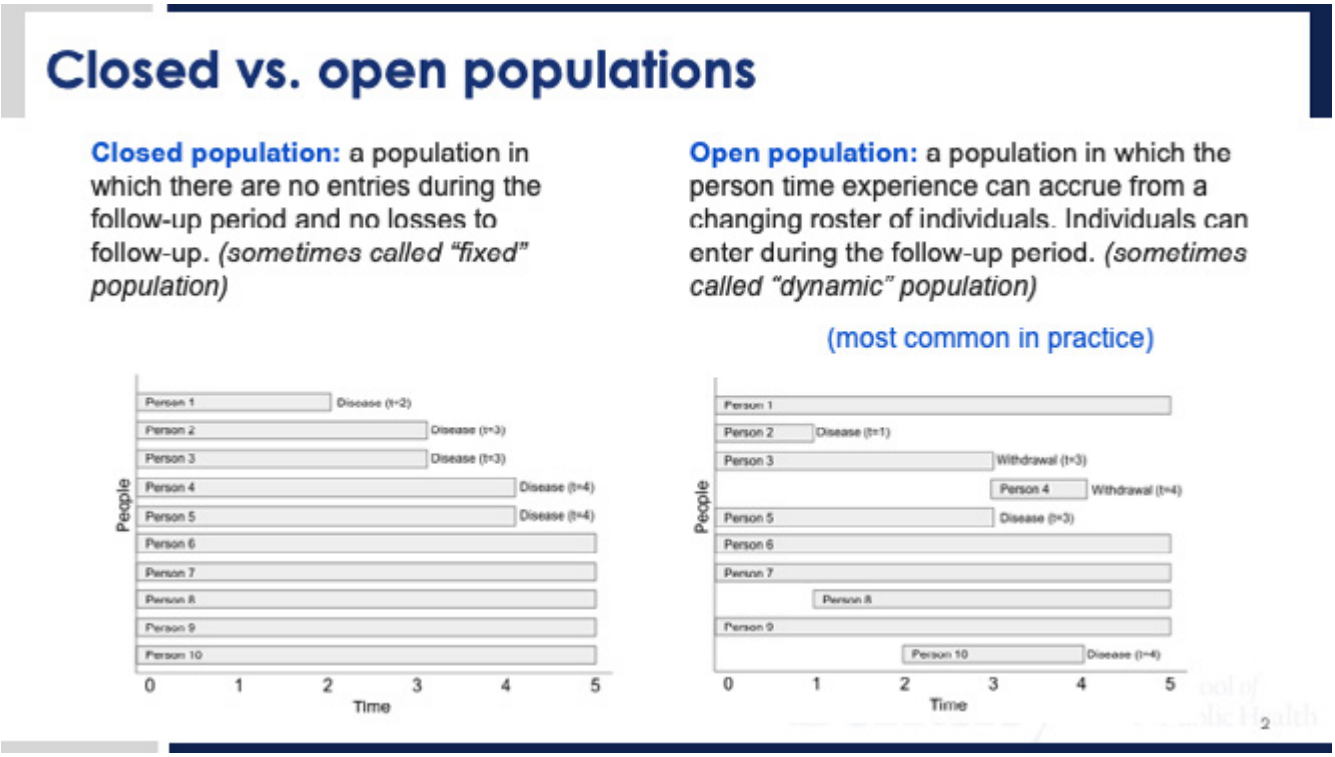
To summarize the key points of this talk today, the age, period, and cohort effects are common in most epidemiologic studies. You should always be on the alert for them. As epidemiologists, we need to think carefully about how we present data-- that is, which line should be connecting which data points, because they can reveal different patterns, as we've seen a couple of times today.

The type of effect that is present may have different implications for the intervention. For example, a strong age effect may suggest targeting a specific age group for intervention. But a strong cohort effect may suggest targeting a specific birth cohort, a group of people born in a different year. And finally, a strong period effect may explain the positive or negative impact of a policy or historical event and influence future policy-making.

These examples are laid out much more fully in your textbook. But hopefully this overview help you to understand the details presented in the text.



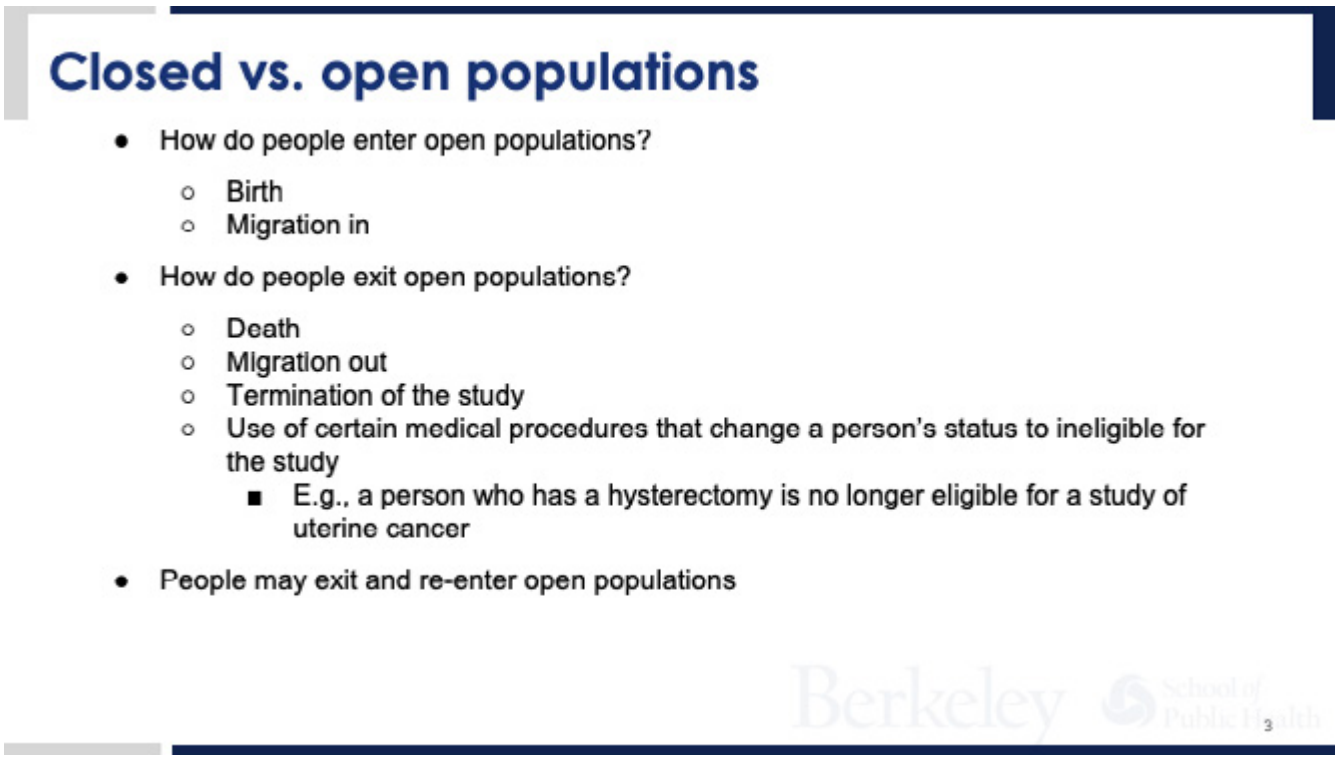
Let's discuss the types of populations that epidemiologists refer to when conducting their research.



Basically, we can think of human populations as mean of two types, closed and open. And what we're trying to demonstrate here in the figures is that in a closed population, you'll see that there are no entries during the follow-up period, and there are no losses to follow-up. That is, this is called a fixed population. So you see that everybody who's in there stays in there until they have disease.

Whereas in an open population, this is a population in which the person time experience of the individuals can accrue from a changing roster of individuals, so the individuals can enter during the follow-up period. Sometimes this open population is referred to as a dynamic population, and it's the most common type we see in practice in epidemiology.

So notice in the figure here, Person 1 was present during the entire period of follow-up. Person 2 developed disease. Person 3 withdrew. Person 4 started later than Persons 1, 2, and 3, and then withdrew before the end of the follow-up period, and so on and so forth. So people could begin at the beginning of the follow-up period, or they could enter sometime during the middle of the follow-up period. People could live until-- live without disease until the end of the follow-up period, or they could develop disease during the follow-up period.



So it's worth asking how people both enter and exit open populations. So the two main ways that people enter is they're born, and if follow-up then begins at birth, that would be one way to enter a population. Or they physically migrate in to the study. They come to the attention of the investigators from having moved into eligibility for the study.

And then there are multiple ways people can exit or leave open populations. They could die. They could physically migrate out. The study could end, or they could have a certain medical procedure that changes their status to ineligible for the study. So they were eligible, and then they became ineligible because of a medical procedure. So for example, a woman who has a hysterectomy is no longer eligible for a study of uterine cancer.

And note that people may both exit and re-enter open populations. So someone who leaves an open population could re-enter that open population.

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 population** because new drug users may be added to the population over time.

Berkeley School of Public Health

4

Now, whether we define a population as open or closed depends on the time scale we use to define that population. So for example, let's say we're talking about all people who ever used a specific drug. If the time scale is defined as the time when a person started using a drug up until the time they stopped using a drug, this is a closed population. But if the time scale is calendar time, this is an open population, because there may be new drug users who are added to the population as time goes by.

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**.

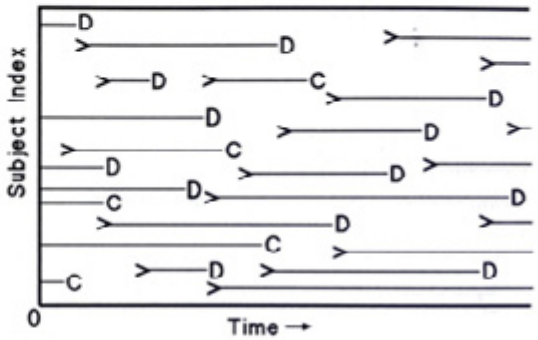


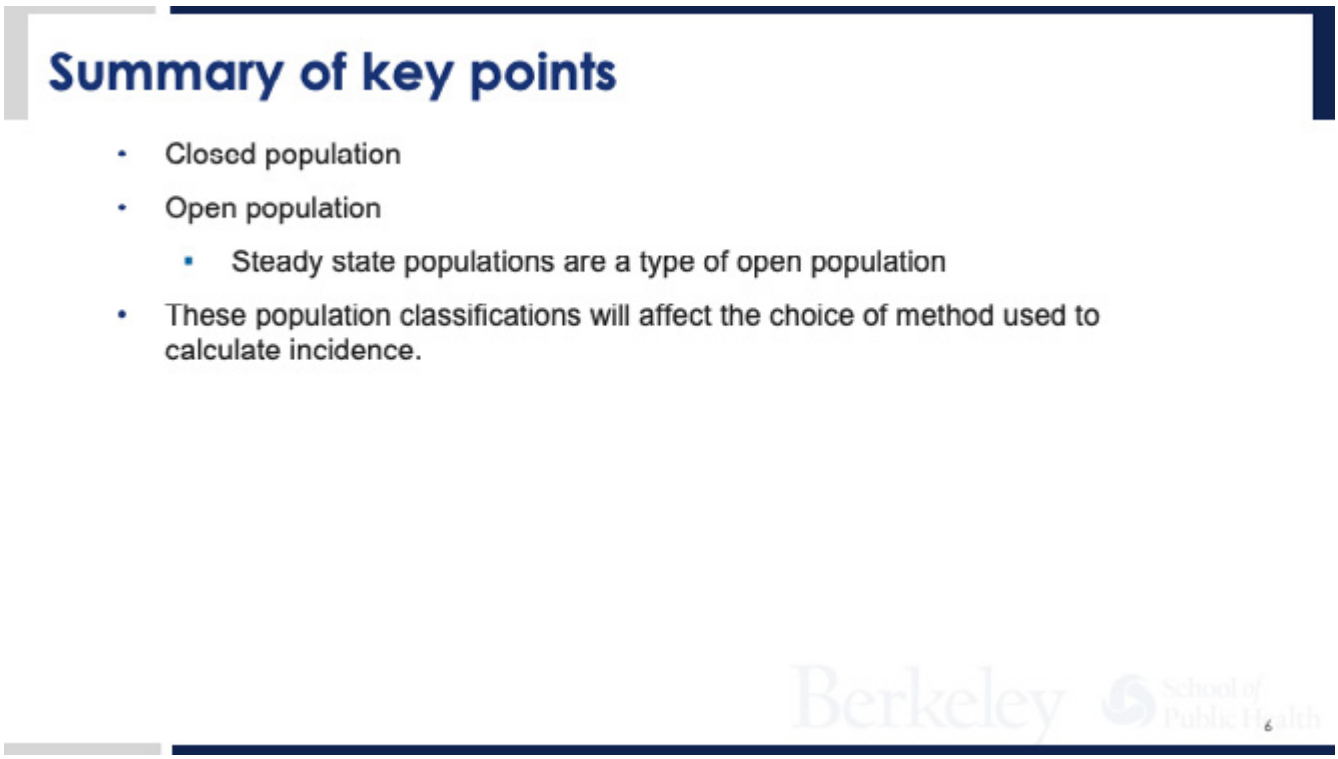
FIGURE 3–3 • Composition of an open population in approximate steady state, by time; > indicates entry into the population, D indicates disease onset, and C indicates exit from the population without disease.

Rothman ME3

An important concept we use when talking about populations is steady state. So when we refer to a steady state population, we're saying this is one in which the number of people entering and exiting is balanced within some period of time across age, sex, and other factors that might affect risk.

And this property is only relevant to open population. So in this Figure 3-3 here, what we're seeing is there are some people in the population at the beginning of the study who then develop disease. That's what the disease, the letter D means.

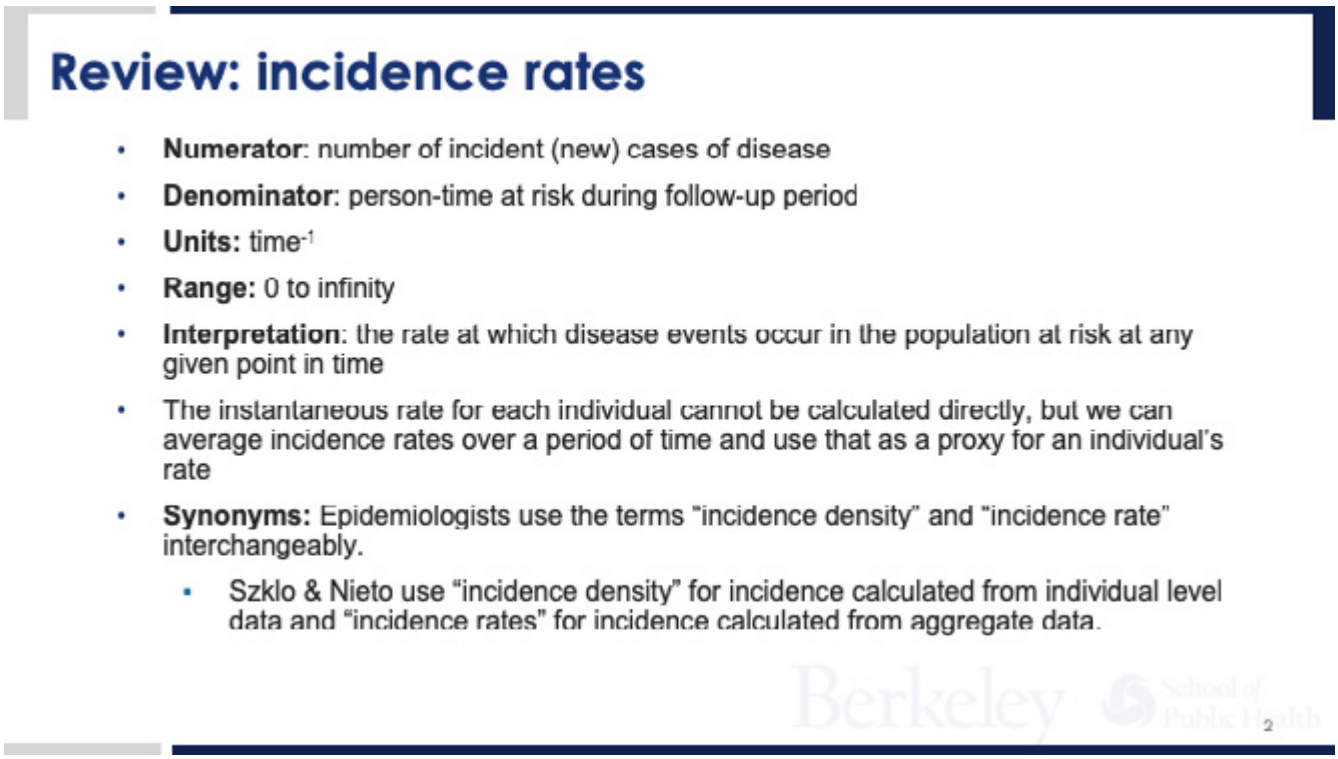
Some people start in the middle of the observation period, and then they exit from the population. And the letter C there is used to refer to people who are exiting the population. This letter C refers to a concept called sensory that's not shown in the figure here, but that's the different reasons that we discussed on an earlier slide, why people may leave a population. So just notice some people are entering and exiting at very different times in this open population that is at steady state. So steady state, kind of is a sense it's balancing the people entering and the people leaving.



So in summary, we've talked about population types, specifically closed populations, open populations, and that steady state populations. That's a term we refer to a type of open population, where the exiting and entering is balanced. These population classifications will affect the choice of method we use to calculate incidence, that is, the measurement of disease or outcomes in these studies.

Lecture 3.2.1: Calculating Incidence Density and Incidence Rates

Today we're going to review the calculation of incidence density and incidence rates.



So just as a very quick review from prior discussions, let's remember that the numerator for t his measure is the number of incident, or new, cases of disease. The denominator is the person-time at risk during the follow-up period. And we'll review again what person-time means.

The units are reciprocal time. That is 1 over time. This measure can range from zero to infinity. And that's because at its lowest value with no incident cases in the numerator, the fraction would reduce to a zero. And if the denominator is zero, no person-time of follow-up, then the measure would be infinite.

The interpretation of this measure is the rate at which disease events occur in the population at risk at any given point in time. This implies an instantaneous rate for each individual. But actually, the instantaneous rate can't 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.

There are some synonyms for these terms. They are incidence density and incidence rate. They're used interchangeably often in text. So for example, Szklo and Nieto use the term "incidence density" for calculated from individual level data. And they use the term "incidence rate" for incidence calculated from aggregate data in a population.

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

Incidence rates can be calculated for closed or open populations. And remember, closed population means nobody leaves or enters, whereas an open population means people can leave and/or people can enter. Study participants can be followed for different amounts of time. These don't distinguish between people who never developed 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

Incidence rates can be expressed with different units of time. So all of the following are equivalent, but they're using different units of time. So I could say 0.024 per person-day. Or converting to 100 person-days, I could say 2.4 per 100 person-days. Or I could say 8.76 per person-year.

Notice again this concept of reciprocal time. That's the per person-year or per 100 person-day or per person-day-- means we have the number over the unit in the denominator. So person-years are commonly used when the disease is rare. And 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
Lung cancer cases	Death	Person-months
Influenza epidemic	Incident influenza	Person-weeks
Children with acute diarrhea	Recovery	Person-days

Szklo 3rd ed.

Let's look at some examples of different time units that can be used for incidence rates in epidemiology studies. And in this table, we're looking at person-time units that might be used according to the frequency of events under investigation. So the point here, of course, is that we might choose to use different person-time units, depending on the type of study we're doing. Let's just look at a couple of these.

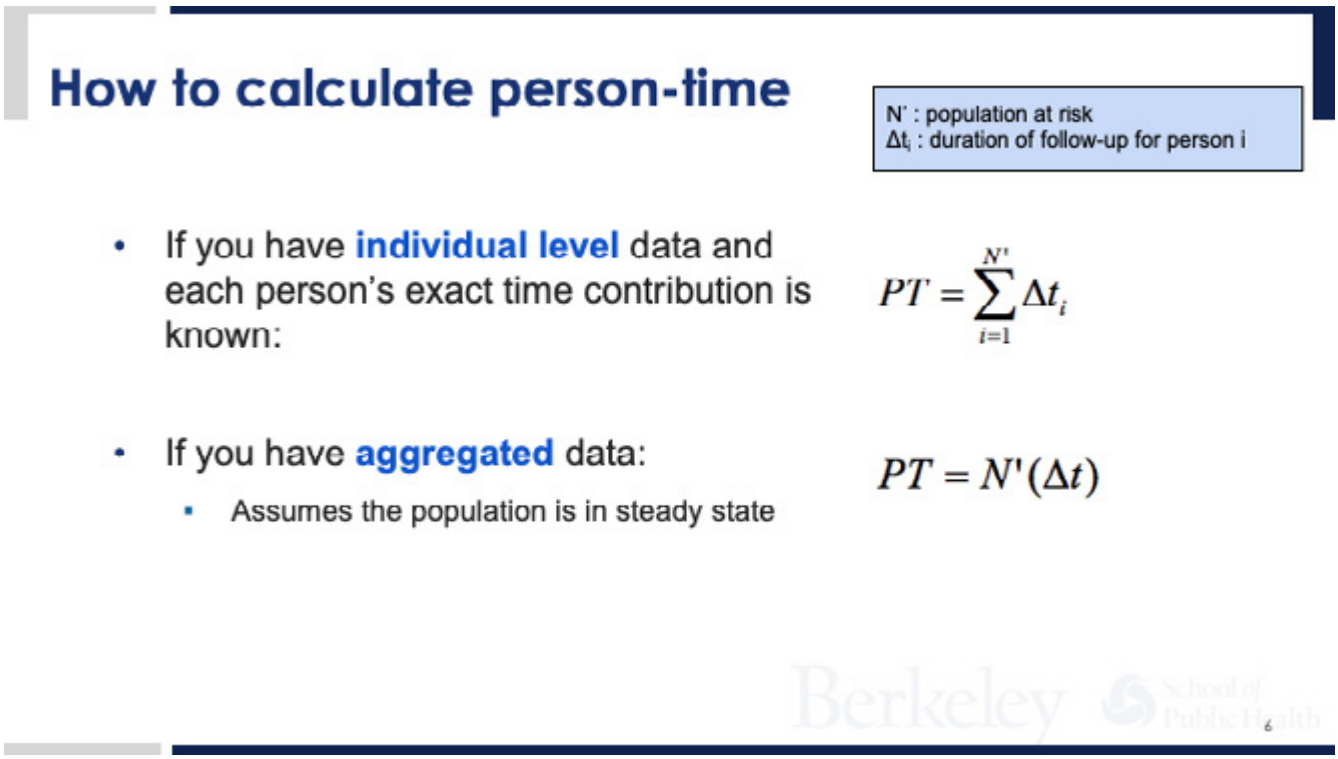
So in general, if, let's say we're looking at some general population out somewhere in California or in the United States or anywhere and we're studying incident cases of breast cancer, that is, new cases of breast cancer, typically we would use person-years to do such a study. It's not that other units are wrong. It's just that person-years kind of captures the right tempo of what's happening in a study such as this.

How about if we were looking at malnourished children and we were looking at cases of new diarrhea or incident diarrhea? Well there, it wouldn't make sense. It wouldn't feel right to use years. That's too long a period generally given the pace at which diarrhea occurs. So here we might look more commonly at person-months. It isn't that we couldn't look differently here or also at person-days. But person-months is probably a good tempo at which to measure the occurrence of incident diarrhea in malnourished children.

What if we were looking at a situation where we were studying an influenza epidemic? So new cases or incident cases of influenza is our outcome or our event being studied. Here it might make sense to use person weeks. Given the number of cases that would occur over a period of weeks during the flu season, person-weeks might be a good time unit to use.

And finally, if we were looking at children with acute diarrhea and measuring time to their recovery-- so here the outcome is something good. How long does it take for them to become healthy after acute diarrhea? Well, fortunately diarrhea usually just lasts a matter of days. So we might use person-days here.

So I know sometimes students are confused about choosing these units. And it's just a matter of experience and judgment in terms of the type of disease one is studying to pick the specific unit.



So how do we calculate person-time in an algebraic way? These formulas should be pretty straightforward to read. But I'll just talk them through very quickly. So if each person's exact time contribution is known, then the person-time is going to be the sum of all the individual times of each individual.

So if I have five different individuals and they have five different person-times, I would add up those five different person-times and sum them. This formula is person-time equals that symbol-- the large epsilon, or it looks like an E-- is the grand sum of delta ti, where delta ti is the person-time for each individual. And the i subscript is called the index. And it is indexed from 1 up to the number of people. So that would be from 1 up to 5 people.

If data were collected over regular intervals, we would use a variant of the original formula. And here we would use the number of people at the beginning of that interval-- so N prime 0j. We've seen that before.

And we would subtract off half the number of people who withdrew, because we don't know whether they withdrew towards the end of the interval or the beginning of the interval. And we would again multiply by the delta time, the time of observation for each person. And I think this

formula will make more sense as you start to do problems with it. These were for individual level data calculations.

If we have aggregated data, the formula are similar but a little bit different. So if the population is in steady state, we would calculate person-time as the number of people in our population times the time of observation of the population. If we had 1,000 people observed for two years, for example, that would be 1,000 times 2.

If the population is not in steady state, then we would take the midpoint population. And what's that N prime sub 1/2 refers to the midpoint of the number of people. So let's say at the beginning of one year, we had 1,000 people in the population. And at the end of the year, we had 800 people. The midpoint of that number would be between 1,000 and 800 would be 900. So we would use N prime 1/2 as 900. And then we would multiply by the total time the population is observed again.

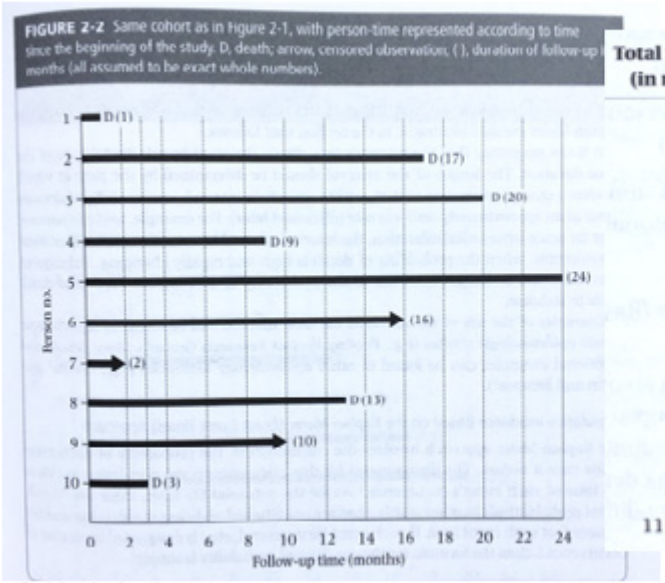
How to calculate incidence rates

- 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

So to calculate incidence rates if we have individual level data, for example, in a cohort study, the incidence rate is going to be the number of events divided by the total person-time of all the individuals in the study. If we had aggregated data, such as we're using census or surveillance data, the incidence rate would be the number of events divided by the average population during the follow-up period.

There are many different ways to arrive at the average. But usually the midpoint of the population from the beginning to the end of the period is what's used. The use of the average population assumes that there's a constant incidence rate across the follow-up period. We're assuming that whatever incidence rate applies to that follow-up period is the same at the beginning, middle, and end.

Incidence rate based on individual level data



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 3rd ed.

So let's look at how to calculate some incidence rates based on individual level data. So here are 10 different people with 10 different exposures. Let's just go through a few of these together.
So person number 1 was observed for one month and then died. So that person contributes one follow-up month. Or if we converted that to person-years by dividing by 12 to convert months into years, we would get 0.083 as the total person-years.

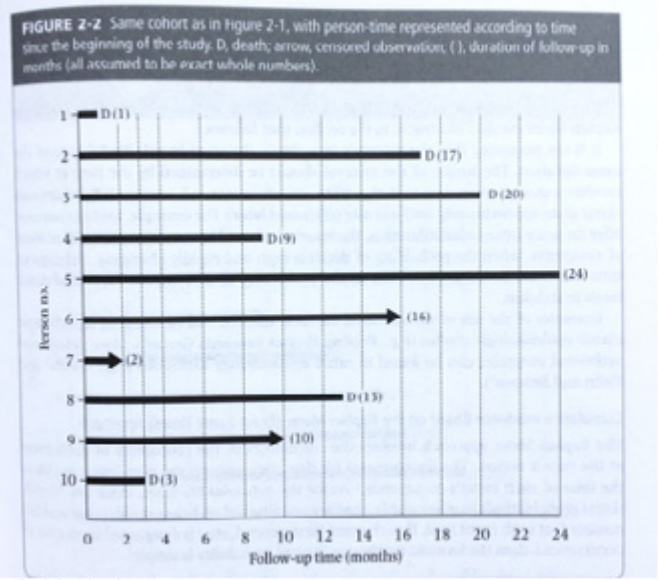
Person number 2-- followed for 17 months, then died. Total person-years here follow-up months would be 17. Total person years would be 17 divided by 12, or 1.417.

Let's do a slightly different one. Let's do person number 7. This person was followed for two months but was censored. We're going to give them two months of credit and convert that into person-years, or 0.167.

Now, how do we calculate the incidence rate for this entire population? Well, first we want in the numerator the total number of events. So that is six events. If you count the different events from person number 1, person number 2, person number 3, person number 4, person number 8, and person number 10, that's six events.

And then the total person-time arrived at by summing up all the individual person-times of 9.583 years, we get 6 divided by 9.583 person-years. Or that could be expressed as 0.63 per person-year. Really you can think of that 0.63 per person-year as 0.63 per one person-year.

Incidence rate based on aggregated data



Szklo 3rd ed.

- 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

Berkeley School of Public Health 9

Now, if we're looking at incidence rates based on aggregated data, here we do the number of events divided by the average population. So the average population is the average of the population at the beginning of follow-up and at the end of the follow-up. Here at the beginning of the follow-up, at time 0, we had 10 people. And at the end of the follow-up, we had only one person.

So the average of 10 and 1 would be 10 plus 1 divided by 2, or 5.5. The incidence rate then would be 6 events divided by 5.5. So that's equal to 1.09 per person over two years. If you do that division, you get 0.545 per person-year. I divided 1.09 divided by 2.

So you can see that whether we-- if we use the data at an individual level or an aggregated level, we come up with different estimates of the incidence rate. So it's not a shock that these are two different numbers, the prior slide and this number. But be sure you understand why they're different.

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.

10

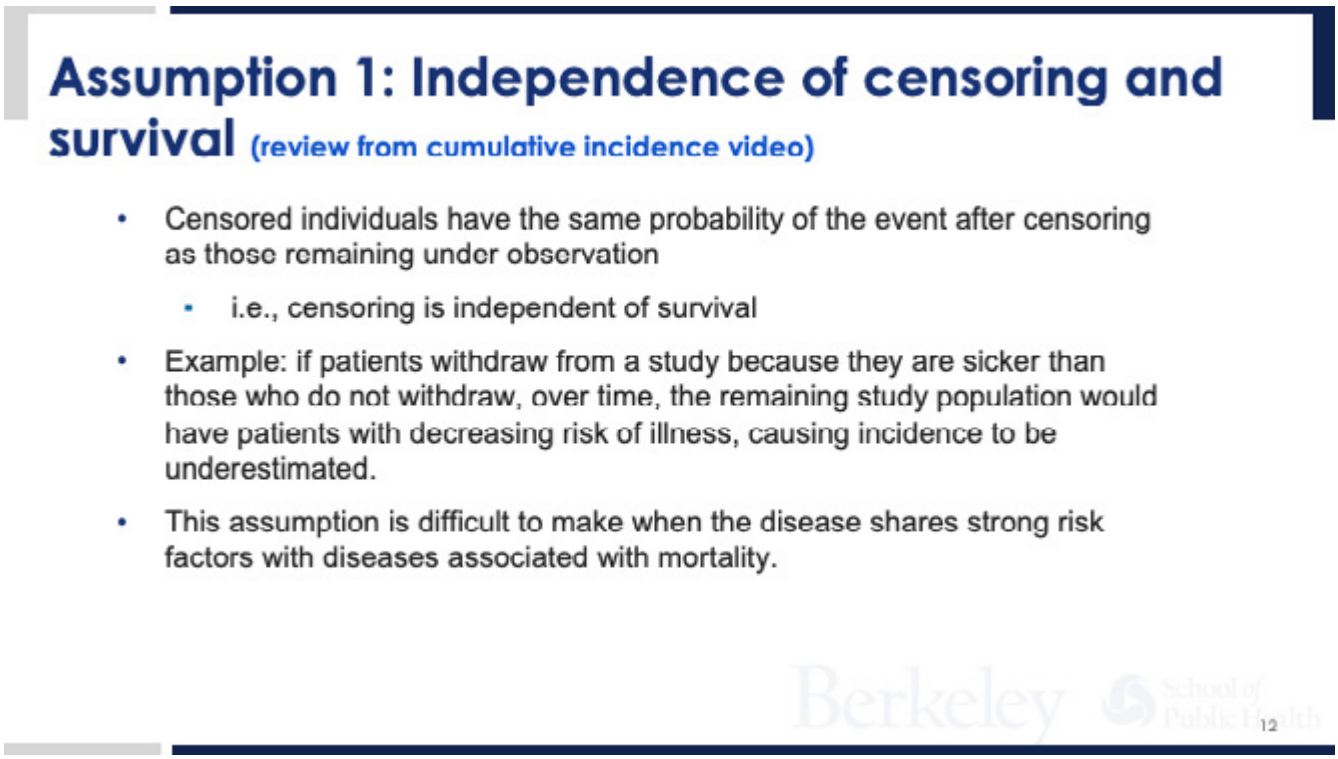
So does incidence rate calculated from individual aggregate data provide the same estimate? We just said no, it doesn't usually as a number. But it does if withdrawals, additions to the population-- for example, births or migration in-- and disease events occur uniformly over time.

And there's a formula given here for an incidence rate using aggregate data to calculate how to convert that to incidence rate using individual data. This is convenient for studies that have to rely on aggregate data. For example, in occupational epidemiology, it's common to estimate age-specific rates using aggregate vital statistics data since individual level data are not always available.

Assumptions when calculating incidence rates

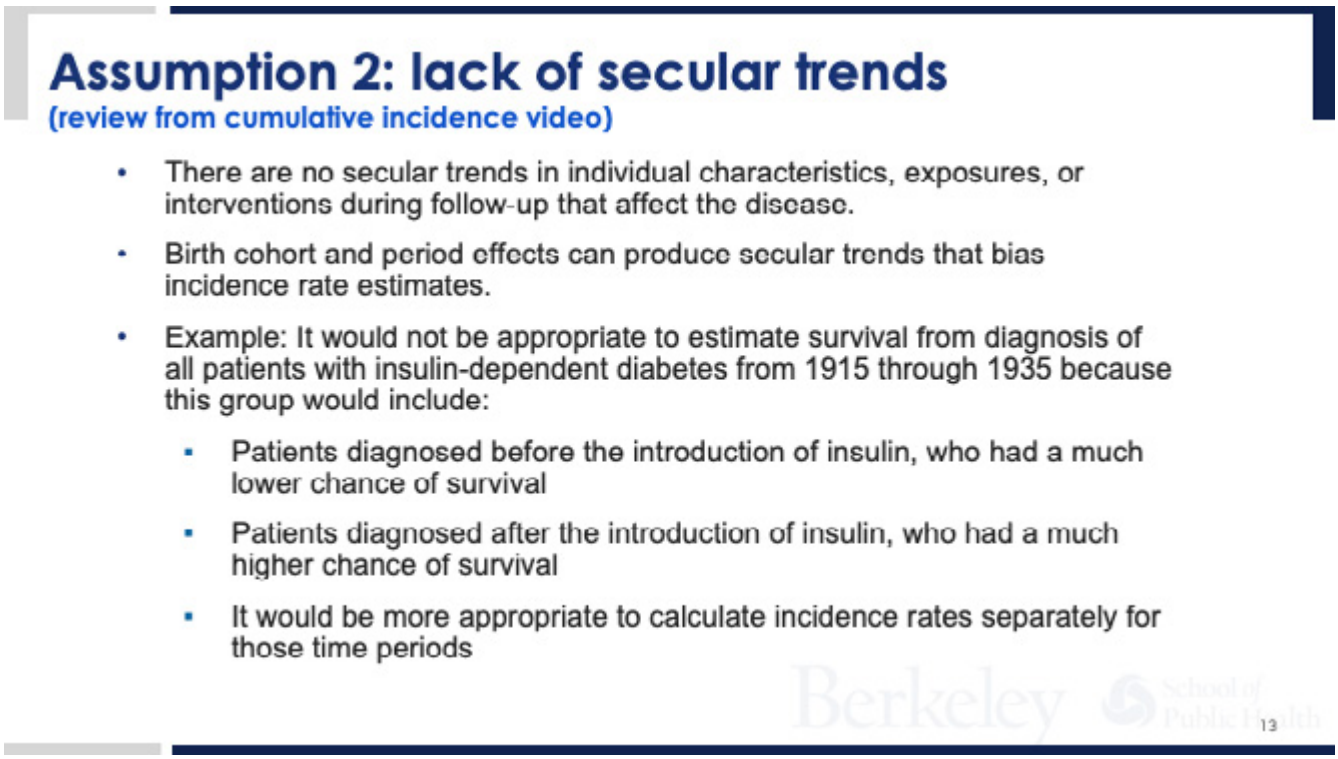
1. **Independence of censoring and survival**
2. **Lack of secular trends**
3. **Risk of the event remains approximately constant** over time during the interval of interest

So there are several assumptions made when calculating incidence rate. And we'll go through each of these. The first is the independence of censoring and survival. Then we'll talk about the lack of secular trends as an assumption. And finally, the risk that the event remains approximately constant over time during the interval of interest.



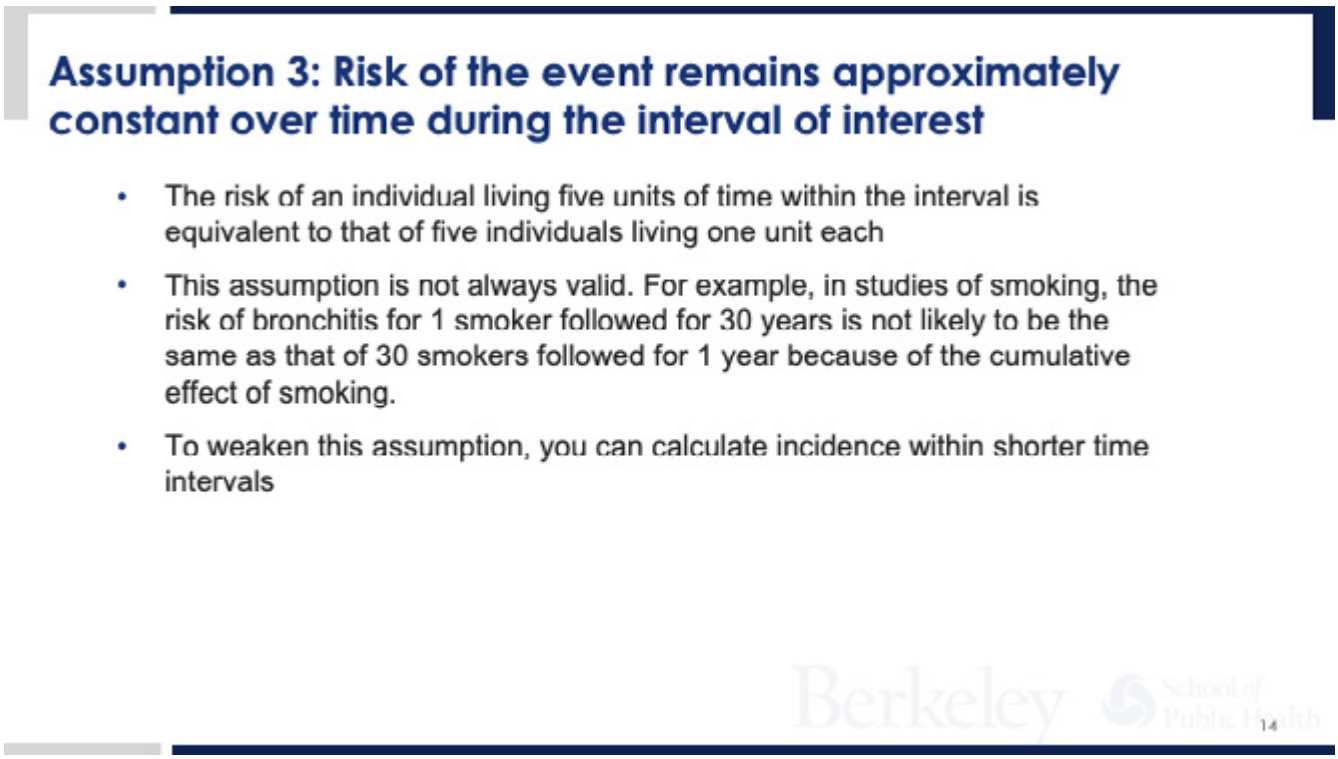
Assumption number 1-- the independence of censoring and survival. So go back and look at the prior video on cumulative incidence where this assumption was talked about as well. So here censored individuals have the same probability in the event after censoring as those remaining under observation. So this is an assumption in that that is that censoring is independent of survival. The people who are dropping out of the study or being lost in the study aren't any different from the people who stay.

So 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 a decreasing risk of illness. That would cause the incidence to be underestimated, because the remaining population would be healthier than what the entire population was when it started. This assumption is difficult to make when the disease shares strong risk factors with diseases associated with mortality.



The second assumption here is that there are no secular trends in individual characteristics, exposures, or interventions during follow-up that affect the disease. And what I mean by secular trends are changes in the population over time. So birth cohort and period effects can produce secular trends that would bias incident rate estimates. You can go back and study when we reviewed age, period, and cohort effects and how these might bias our estimation of what's going on in a population.

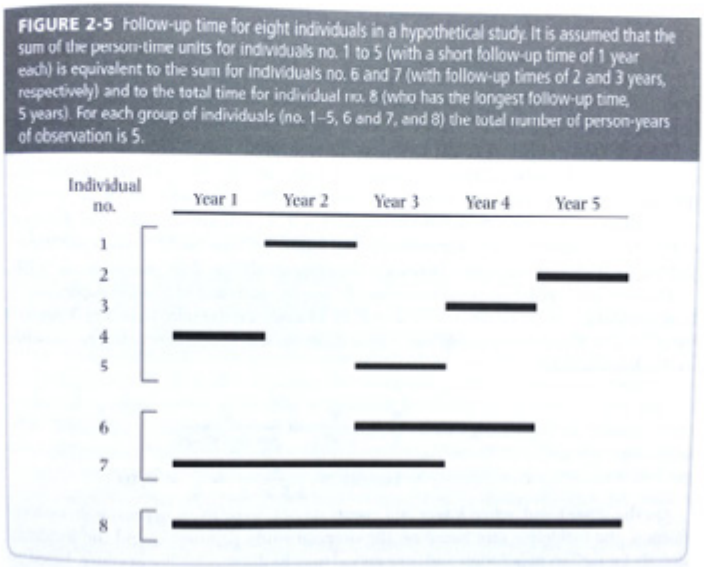
As an example, it would not be appropriate to estimate survival from diagnosis of all patients with insulin-dependent diabetes from 1915 through 1935, because this group would include patients who were diagnosed before the introduction of insulin who had a much lower chance of survival, along with patients diagnosed after the introduction of insulin who had a much higher chance of survival. So instead it would be more appropriate to calculate incidence rates separately for those time periods.



The third assumption is that the risk of the event remains approximately constant over time during the interval of interest. So the risk of an individual living five units of time within the interval is equivalent to that of five individuals living one unit each. And one thing to note is that this assumption is not always valid.

For example, in studies of smoking, the risk of bronchitis for one smoker followed for 30 years-- that would be 30 person-years-- is not likely to be the same as that of 30 smokers followed for one year, which would also be 30 person-years, because of the cumulative effect of smoking that, in other words, a year of smoking in your 25th year of smoking is likely to be different than a year of smoking in your first year of smoking. So to weaken this assumption, you can calculate incidence within shorter time intervals. So for example, for the smokers, you might calculate the incidence within each of the 30 years that you follow them.

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

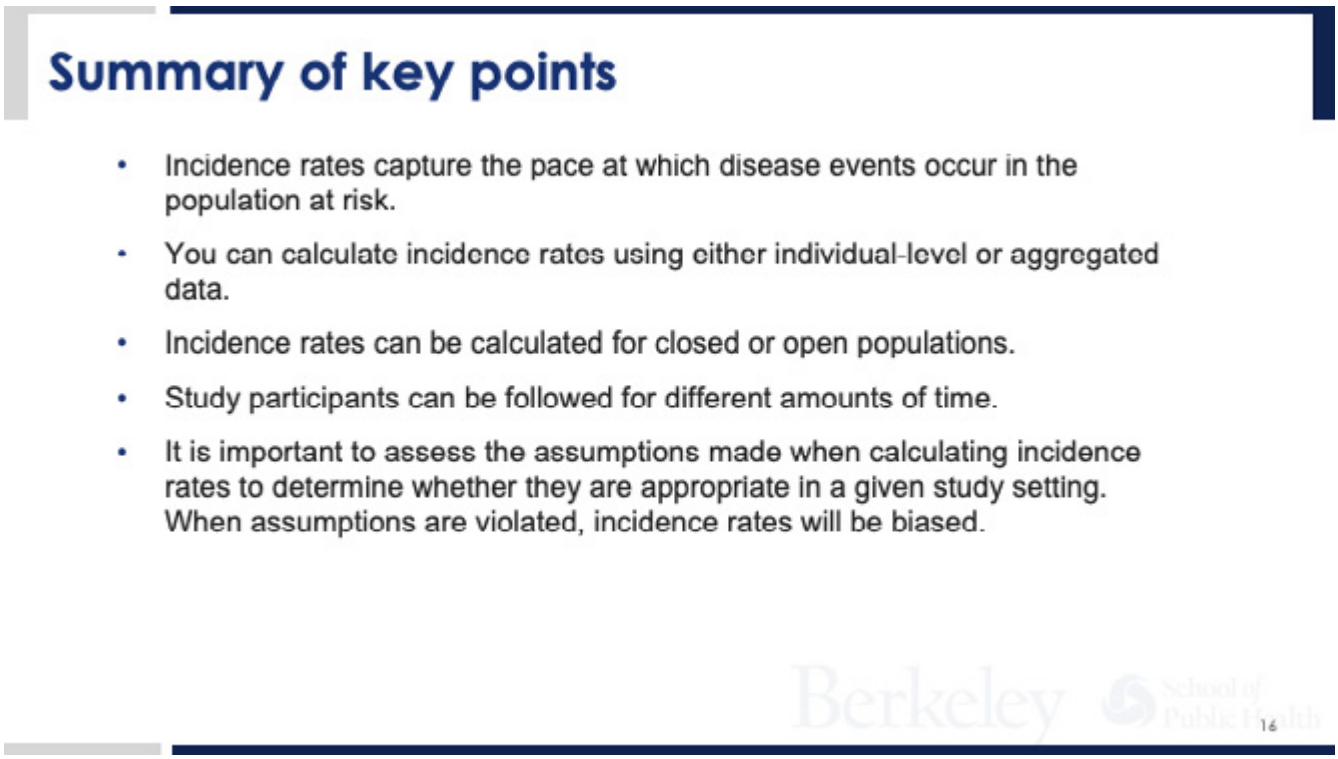


Szklo 3rd ed.

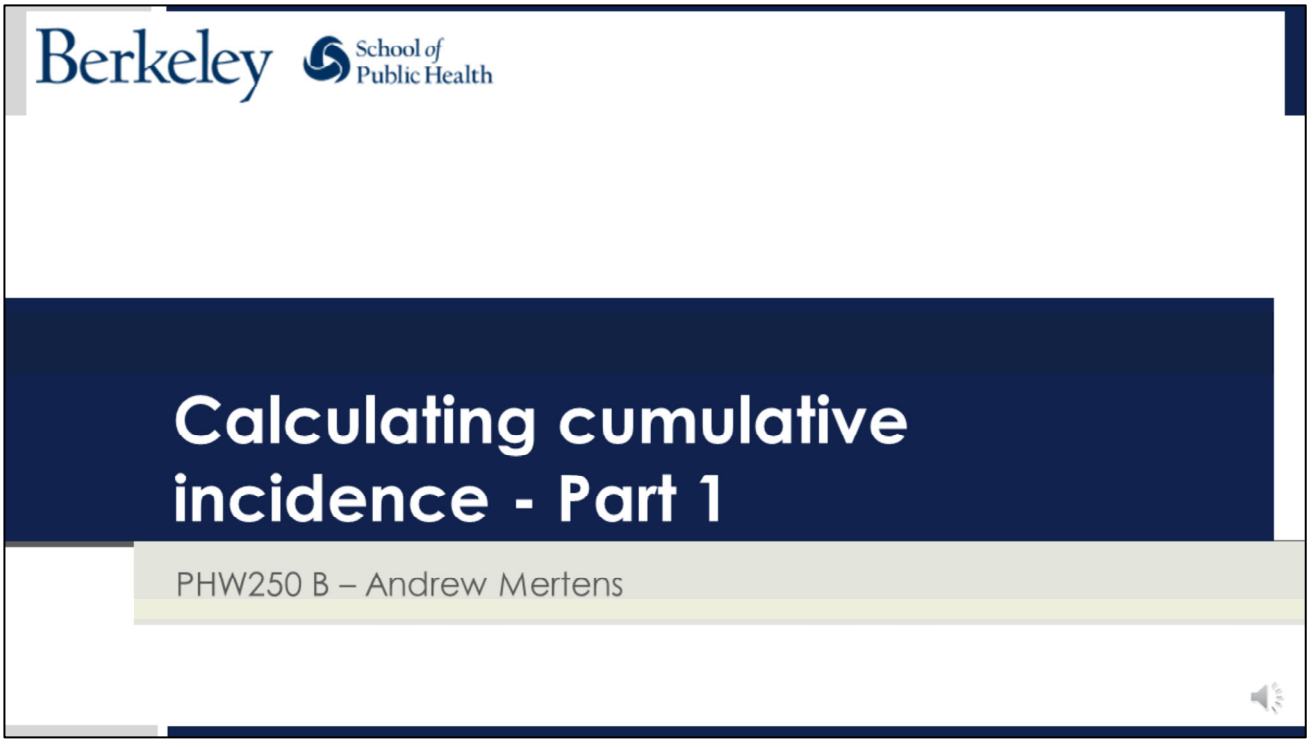
15

Let's look at a graphical representation of this assumption of constant risk of a follow-up period. The assumption made when using the incidence rate is that the following would be the same. Look at the first five individuals. They're each followed for one person-year. That would be five person-years of follow-up.

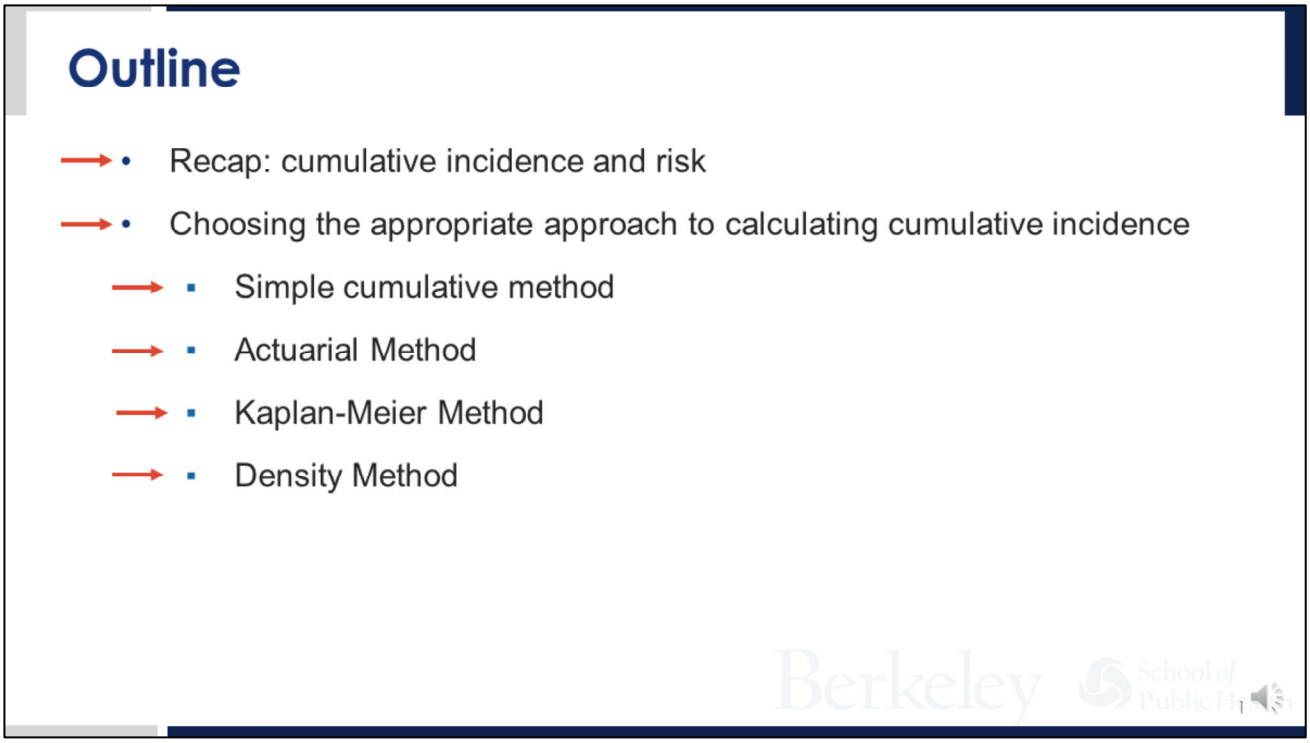
The second two individuals are followed for three years and two years respectively. So that would be five years of follow-up. And the third person is followed for five years of follow-up. So each of these three groupings here has five person-years of follow-up. But I think it's clear to you why they could be very different from each other.



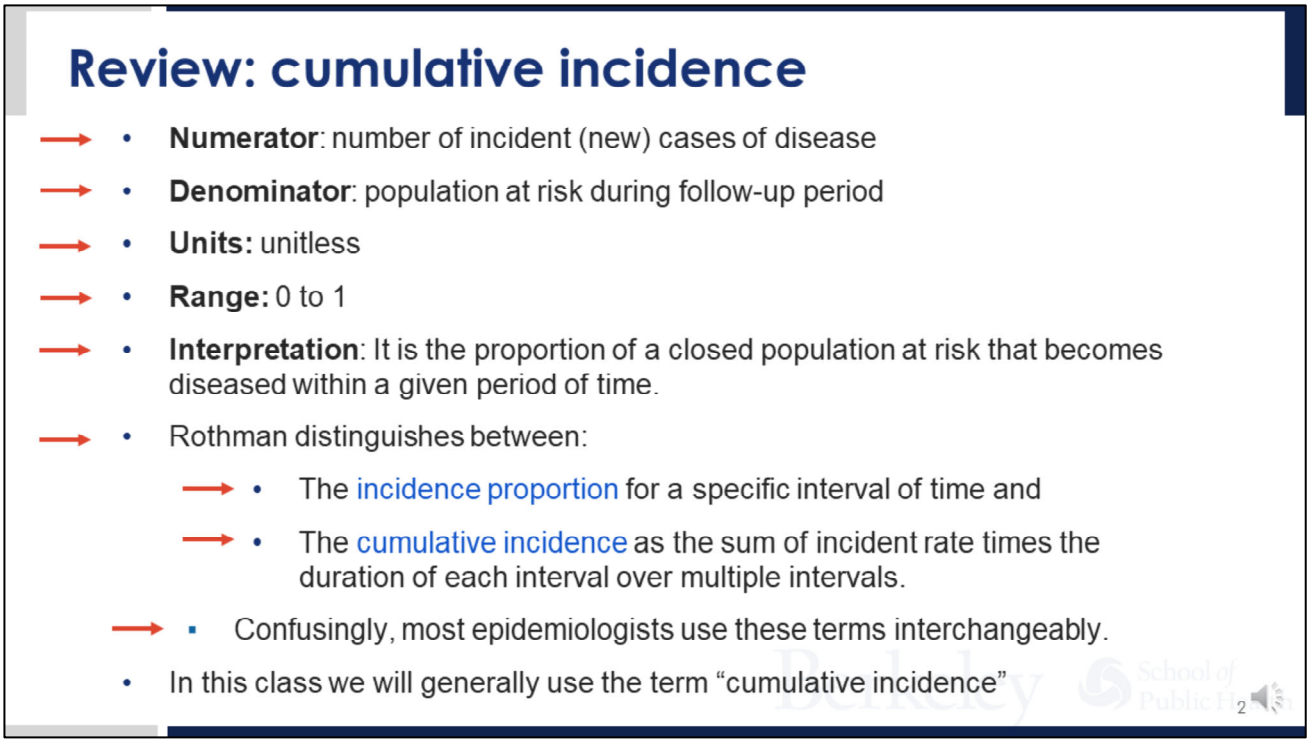
The summary of the key points for this talk-- incidence rates capture the pace at which these events occur in the population at risk; you can calculate incidence rates using either individual-level or aggregated data. Incident rates can be calculated for closed or open populations; study participants can be followed for different amounts of time. And it's important to assess these assumptions made when calculating incidence rates to determine whether they're appropriate in a given study setting. When the assumptions are violated, incidence rates will be biased.



In this video, I'll talk about how to calculate cumulative incidence, and this is the first of two videos on this topic.



Here's our outline. I'll start by recapping the concepts of cumulative incidence and risk and how they differ, and then we'll talk about four different approaches to calculating cumulative incidence-- the simple cumulative method, actuarial method, Kaplan-Meier method, and density method.

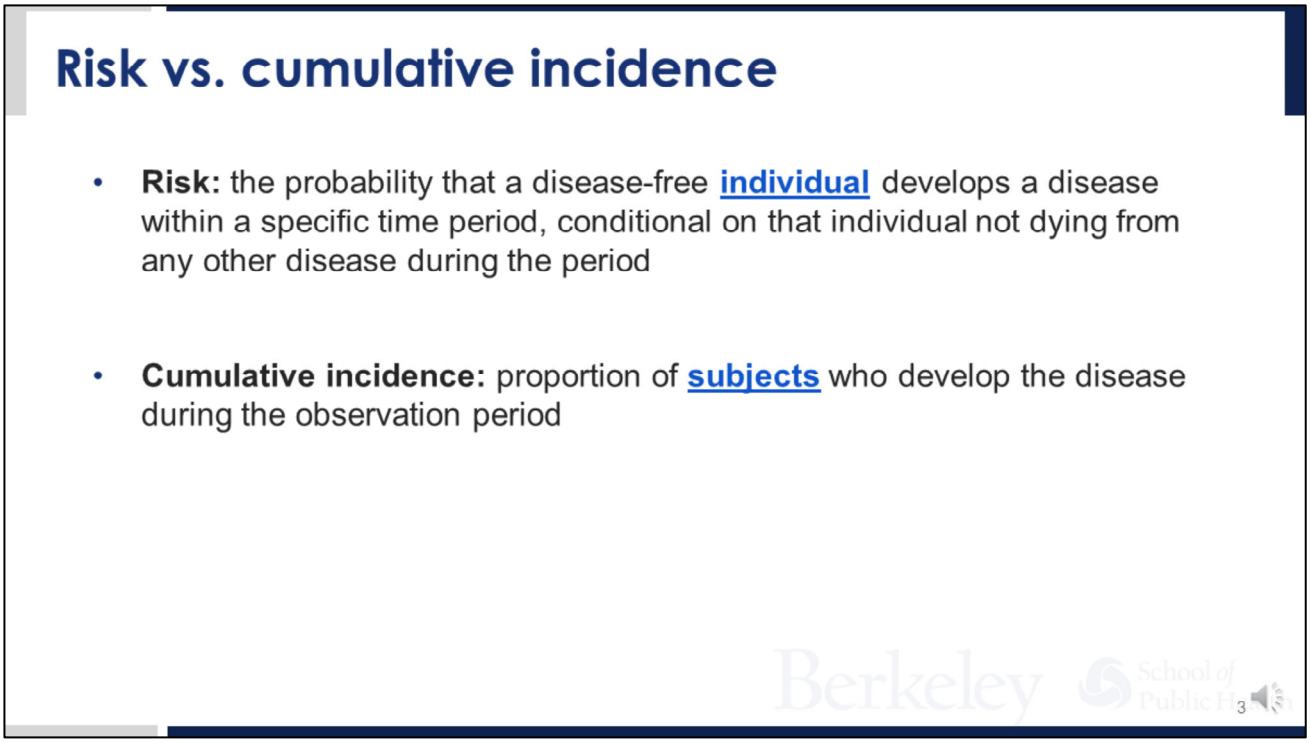


First, let's review cumulative incidence. In *the numerator, we have the number of incident cases of disease. So this is disease onsets or new episodes of disease, and in *the denominator, we have the population at risk during a follow-up period. And when we say population at risk, we mean people who are able to develop the disease.

*Cumulative incidence is dividing people by people, and so it's unitless. It's a probability or proportion, and it *ranges from 0 to 1. We can *interpret it as the proportion of a closed population at risk that becomes diseased within a given period of time. We're going to talk a bit more about closed versus open populations, but this is the strictest interpretation here, is that we're assuming we're talking about a closed population.

*Rothman distinguishes between two different terms-- the incidence proportion and the cumulative incidence. So *Rothman says that the incidence proportion is the proportion of people who newly develop a disease among those at risk for a specific interval of time. *And then he says that the cumulative incidence is the sum of the incidence rate times the duration of each interval over multiple intervals.

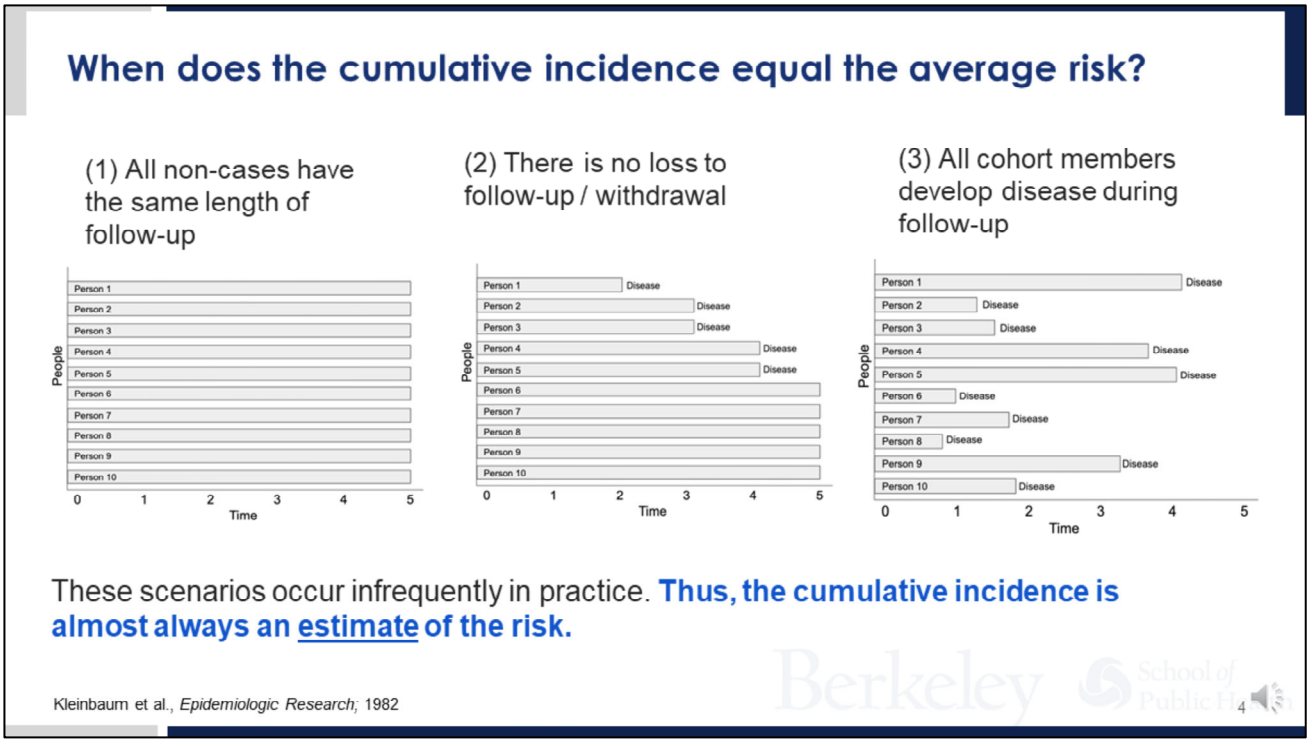
This is a bit of a nuanced issue, and I just wanted to briefly mention it here to point out that Rothman has this slightly different way of referring to this concept. *It's very confusing, because most epidemiologists use these terms interchangeably. In this class, for the most part, we're just going to use the term cumulative incidence.



Now, what is risk, and how is it different from cumulative incidence? Risk is the probability that a disease-free individual develops a disease within a specific time period, conditional on that person not having died from any other disease during that period. It's very important to remember that risk is, in its essence, an individual-level concept. On the other hand, cumulative incidence is the proportion of subjects who develop a disease during an observation period or a follow-up period.

So cumulative incidence is really thinking about the probability of developing a disease or having disease onset in a population. Whereas, risk is really meant to be interpreted at the individual level, but we use cumulative incidence to analyze our population data to try to make an inference about the risk of disease at the individual level.

And this video is talking about how we can estimate cumulative incidence. What we really want to do is estimate risk, but for a variety of nuanced reasons, we're not going to directly estimate risk. We will make some sets of assumptions that allow us to feel pretty good about saying that our cumulative incidence estimates are equivalent to risk in certain settings.



So let's talk about when we can say that the cumulative incidence equals the average risk. And the reason I'm saying average here is that, again, cumulative incidence is really a population-level concept, and risk is an individual concept. And so if we want them to be equal, it's the average risk of the individual risks that would be equal to the cumulative incidence.

Now, there's three scenarios where these two quantities, cumulative incidence and average risk, could be equal. The first is when all non-cases-- so these are people who did not develop the disease-- had the same length of follow up.

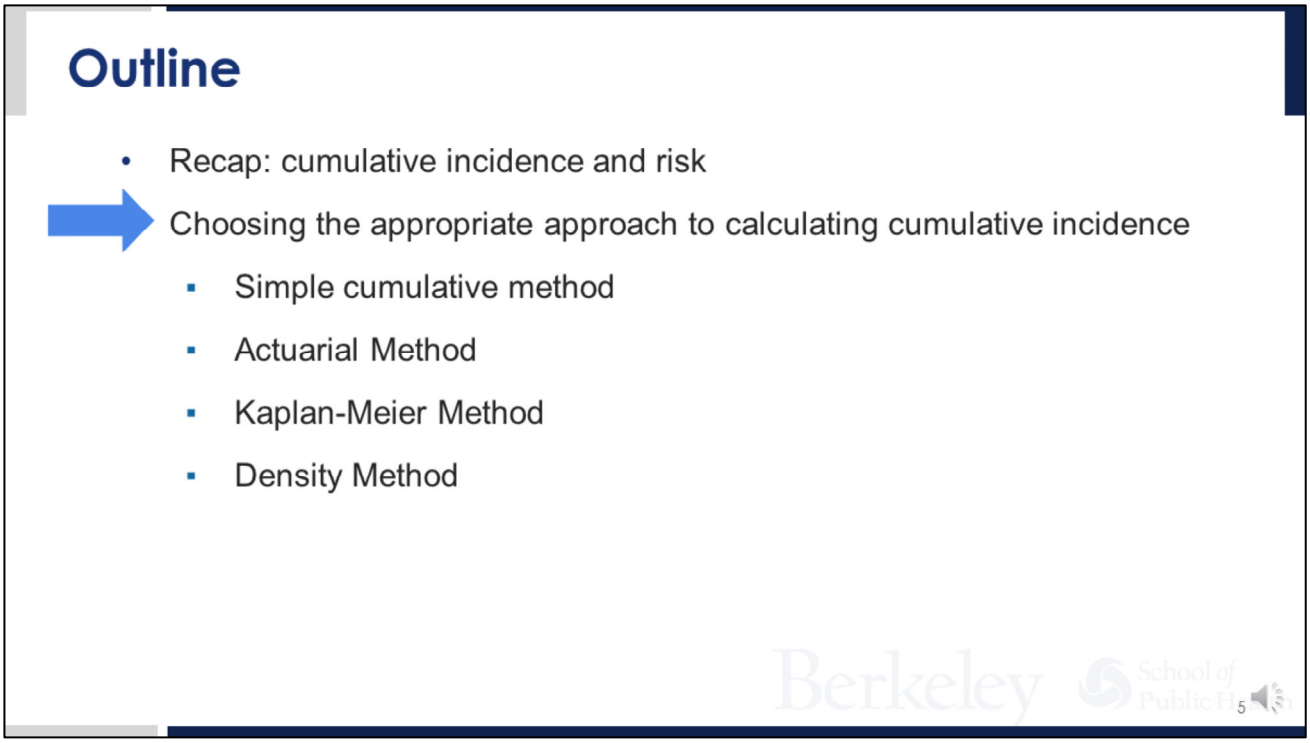
So if we look at this figure to the left here (Fig. 1), on the y-axis we have individual people in our study population that we're following over time. On the x-axis, we have our follow-up time. There's 10 people in this little population of interest, and what the gray bars indicate is that they remained disease-free, and they did not die of any other causes for the entire let's say it's years, five years of the study. So every single person was a non-case, because nobody has the word disease on their row, and they all have the same length of follow up. So if this is true, then the cumulative incidence equals the average risk.

Second example, there is no loss to follow up or withdrawal. So in this second figure, we see that the first five people develop disease at different times. But then the remaining five people, persons 6 through 10, didn't develop the disease, and we have complete follow up on them from time 0 all the way to time 5. In this situation, we can

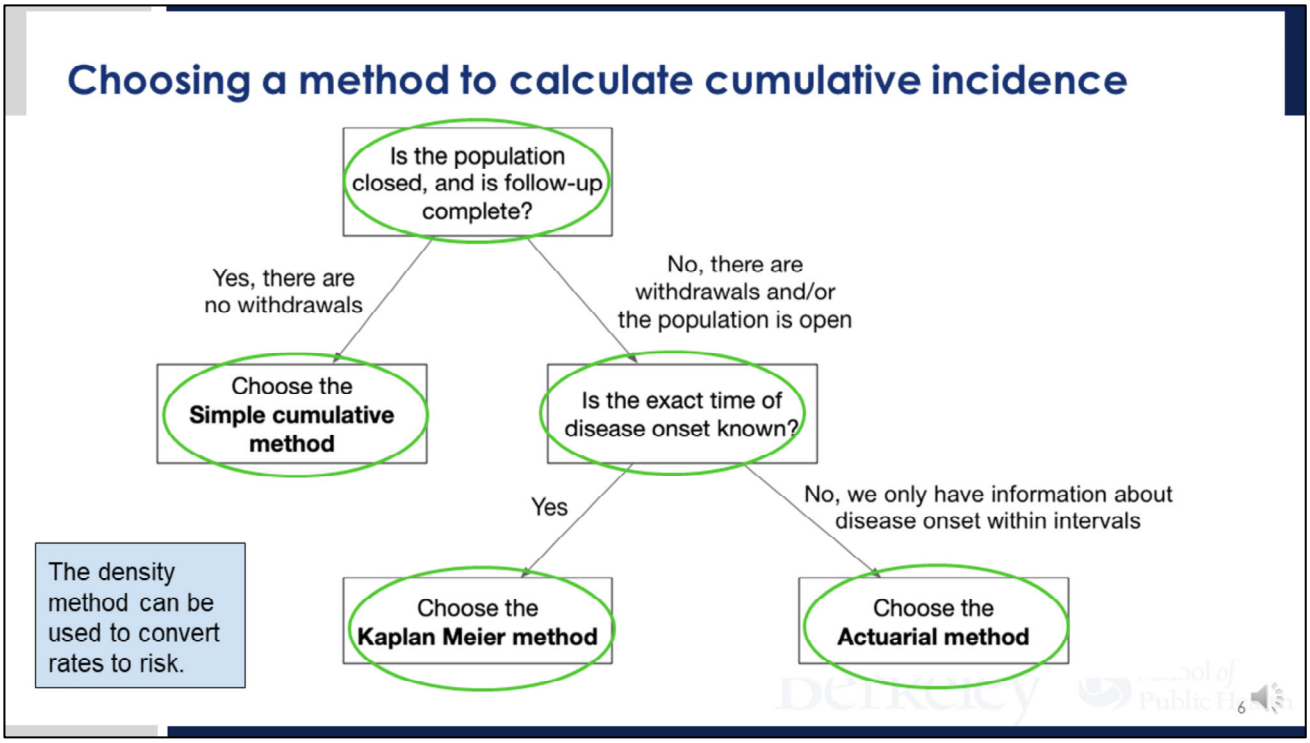
reasonably say that the cumulative incidence is equal to the average risk.

The third scenario, all cohort members develop the disease during the follow-up period. So in this plot on the far right, every single person developed the disease, and so in this case, we have basically a cumulative incidence of 1. Because it's 10 people develop the disease divided by 10 people at risk at the beginning of the follow up. That's equal to 1. Average risk is also going to be equal to 1, or 100% probability.

Now, the thing is, these three scenarios are actually quite infrequent in practice. And so even though there are these particular situations, where the cumulative incidence equals the average risk, they're so rare that, for the most part, it's really better to remember that the cumulative incidence is almost always an estimate of the risk. It's not going to equal the risk, but it may estimate it.



All right. Now, let's talk about choosing our appropriate approach to calculating cumulative incidence.



So here's a decision tree. Let's start at the top. First, we have to ask ourselves, are we dealing with a closed population with complete follow up? And if the answer is yes, and there's no withdrawals, then we choose the simple cumulative method.

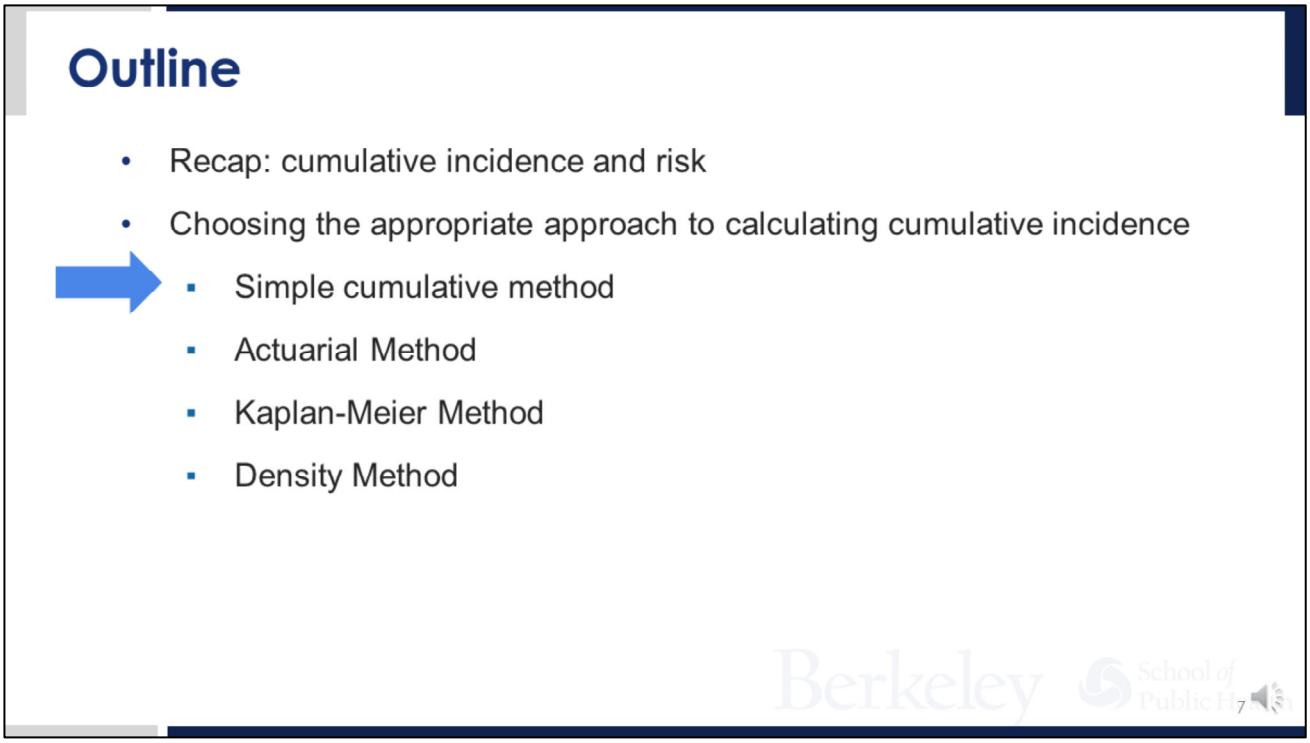
So this is a pretty restrictive scenario. An example that we commonly give would be if you were looking at an outbreak investigation, perhaps for food poisoning. Maybe there was a large family reunion, and somebody brought an egg salad that went bad in the hot summer sun. Some people got sick, and you wanted to do a little epidemiologic outbreak investigation to try to figure out that it was the egg salad, because you weren't sure what it was.

That could be a closed population, because everybody who was at the reunion. You have data on the follow up, because the follow up period is very short between the onset of disease and the exposure. So maybe in that situation it's appropriate to choose the simple cumulative method, but most of the time, that's not going to be the case.

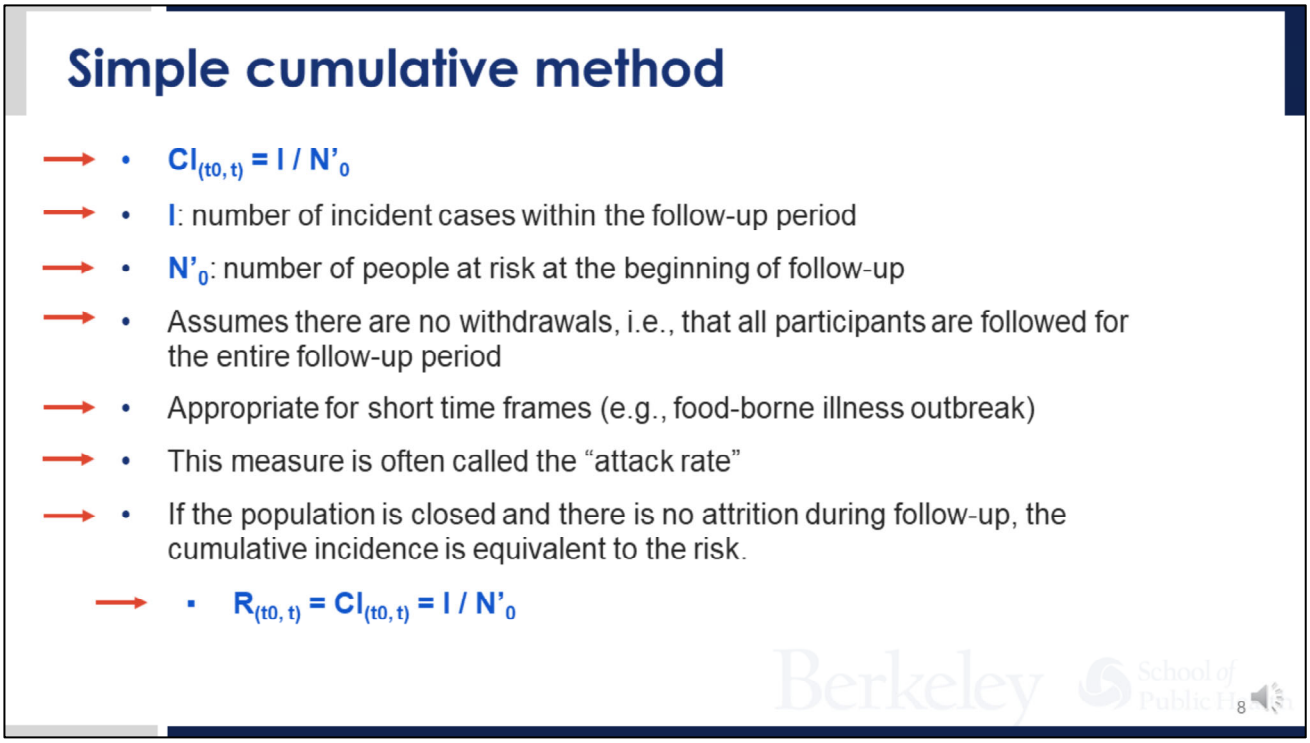
So most of the time, we'll say, when we answer this question, “Is the population closed, and is follow up complete”, we'll say no. Either the population is open, or there were some withdrawals.

*And then we need to ask ourselves, is the exact time of disease onset known? And if the answer is yes, we'll use something called the *Kaplan-Meier method.

*If the answer is no, and we only have information about disease onset within intervals, we'll choose the actuarial method. The fourth method, the density method, doesn't really fit into this tree, because we use that method when we want to convert rates to risk. So we'll come to that in the next video.



All right, let's talk about how to calculate the cumulative incidence using the simple cumulative method.



*So here's the formula. I'm going to go over the notation on the next slide, so let's hold off on that for now. But what we can see here is we have the cumulative incidence is equal to I over N prime sub 0.

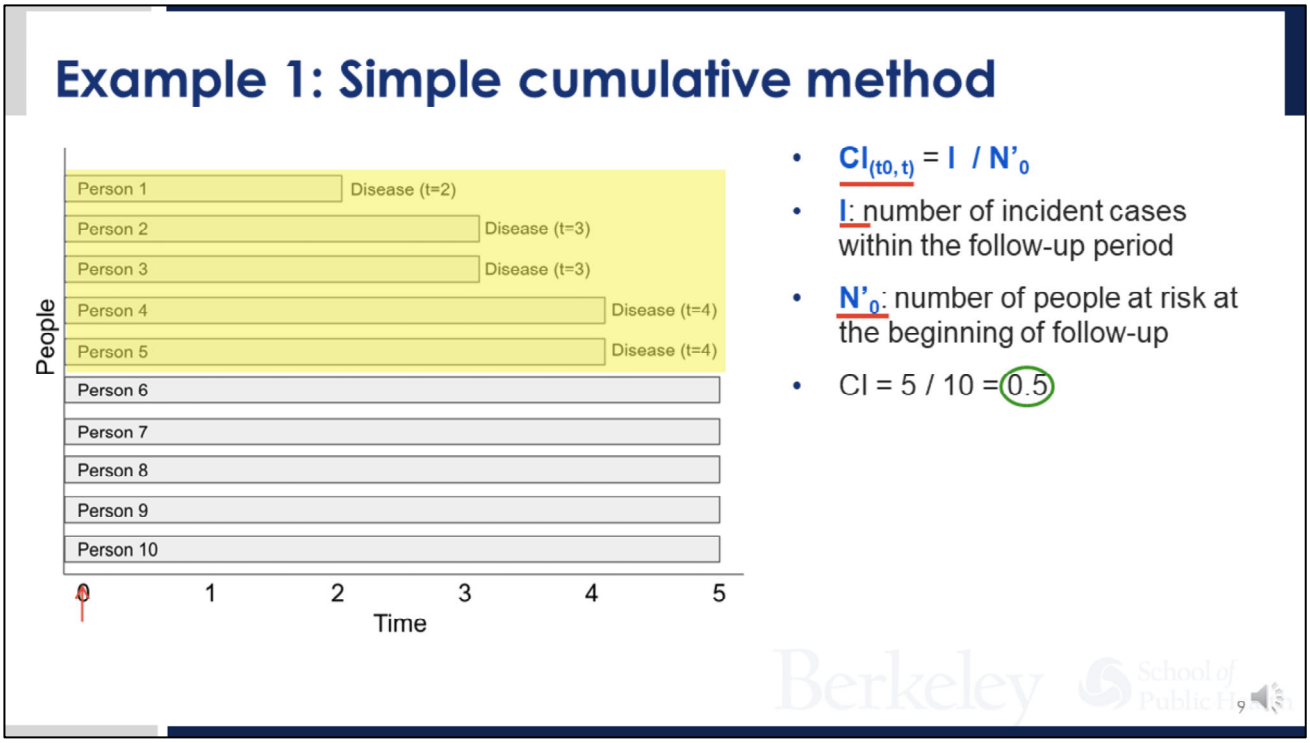
So I is the number of incident cases within the follow-up period. N prime sub 0 is the number of people at risk at the beginning of follow up, and this-- *as I mentioned in the prior slide with the flow chart-- this calculation assumes no withdrawals. So that means all participants were followed up for the whole period of the study. Another way of saying this, we didn't lose anyone. There was no loss to follow up. There was no censoring. Or There was no attrition. These are all terms that are used for the same concept.

And like I was saying, maybe in the case of a foodborne outbreak, it's reasonable to say that we have complete information on every single person who is at risk at the beginning of the study. But when we start to talk about thousands of people in large geographic areas, it's quite unlikely that this will be true.

*And again, it's best to use this method for a short time frames. This is really often because of this pretty restrictive assumption that there's no withdrawals.

*Sometimes this is called the attack rate instead of the cumulative incidence, and I think that's because, again, this foodborne illness outbreak is a common example. We can think of the disease attacking in this very short period of time, as opposed to

other slower onset chronic diseases, and those are diseases for which we would probably not use this method to calculate incidence. *And then if the population is closed, and there's no attrition, in other words, no withdrawals during follow up, *we can say that the cumulative incidence is equivalent to the average risk.



Here is an example. *So CI sub t0 comma t, what does this mean? Well, the first quantity in the parentheses, t0, indicates that we're interested in starting at time 0. So this is right here on the x-axis. That's the beginning of our study.

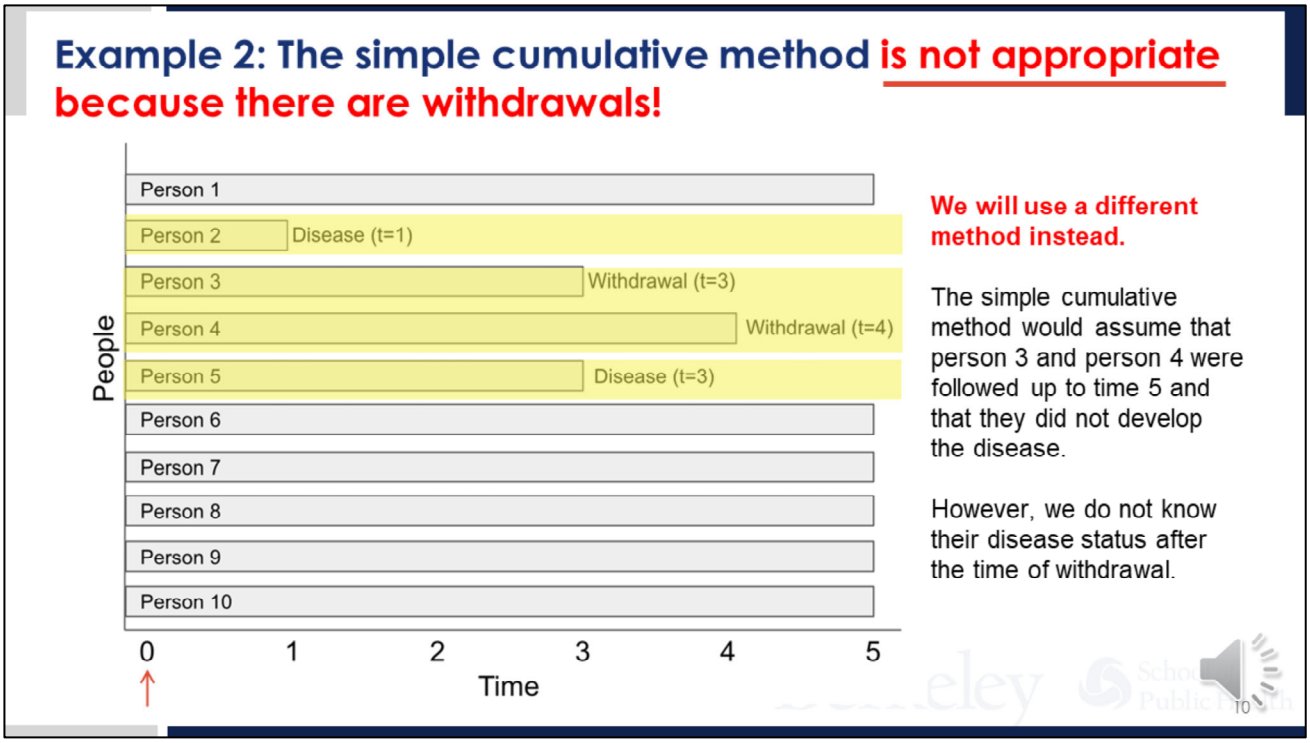
And the second one, t, indicates that we want to follow people up to a certain time. And so we write t here, because it just means that we're referring to really any time period. Maybe we collected data on five years, but we're only interested in looking at time 0 to time 3. That would be written as CI t0 comma 3, but we've written it generally here so that it's flexible.

OK. So we have I is the number of incident cases in the follow-up period, and then N prime sub 0. So N is the number of people, the prime stands for people at risk of developing the disease, and sub 0 stands for at the beginning of the follow-up period. So you're going to see this a lot throughout this video, so get comfortable with that notation.

And now, let's calculate the cumulative incidence using this method. So if we want to get I, we need to count the number of incident cases during this five-year follow up. And there are 5 people, persons 1, 2, 3, 4, and 5 all develop disease at a certain point.

So that goes in our numerator, and then the number of people who were at risk at the beginning of follow up is 10. Because if we go to times 0, right here, what we see is

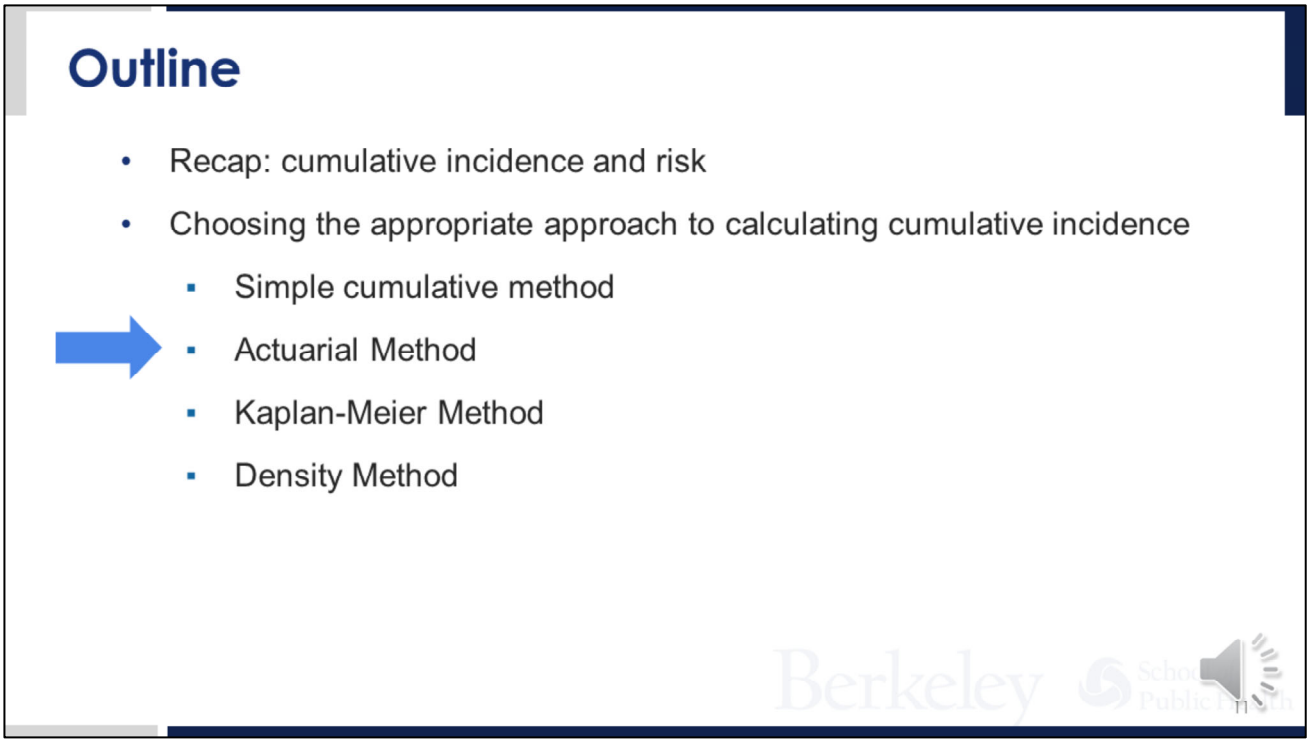
all 10 people were participating in the study at that time. So we say 5 over 10, and we get the cumulative incidence is 0.5.



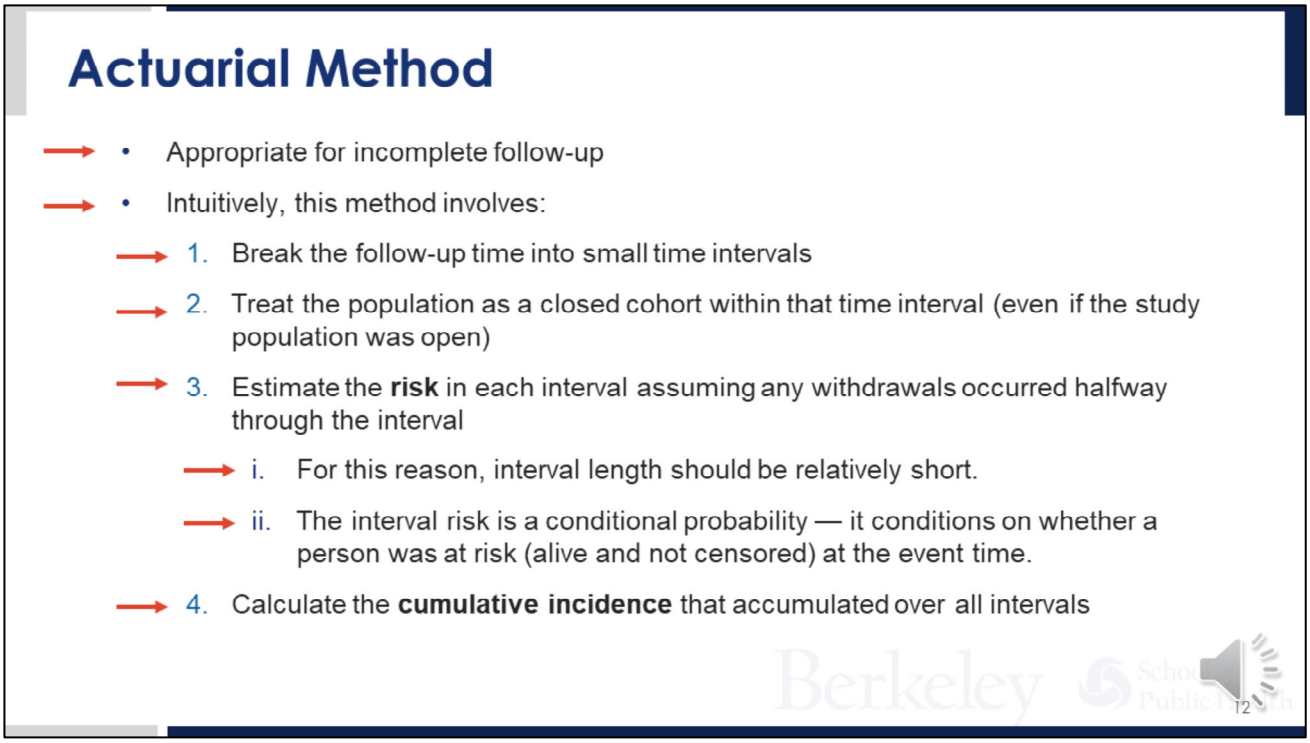
Now, let's do a second example. In this example, we see that person 3 withdrew from the study at time 3, person 4 withdrew from the study at time 4, and then we have 8 other people with different outcomes. *Now, as you can see right here in red, the simple cumulative method is not appropriate for this population, because there are withdrawals. We need to use a different method instead. So why is this? Why is this problematic?

Well, essentially, what's happening with the simple cumulative method is that we're assuming that person 3, who had a withdrawal at time 3, and person 4, who withdrew at time 4, we're assuming that they were actually followed up to time 5, and that they did not develop the disease. Because what we're doing is we're taking the numerator, incident cases of 2, right? Person 2 and person 5 developed disease, and we're dividing by the people who are at risk at the beginning, all 10 people. And we're ignoring the fact that we actually have incomplete information on person 3 and person 4.

So by doing so, we're sort of implicitly assuming that they were there the whole time, and they didn't develop the disease. But in practice, or in truth, we actually don't know what happened to person 3 after time 3, and we don't know what happened to person 4 after time 4. And because of this, this is not a safe assumption to make. Right? So we can't use the simple cumulative method when we have withdrawals.



All right. Now, let's talk about the actuarial method which is one of several methods we can use in this situation.



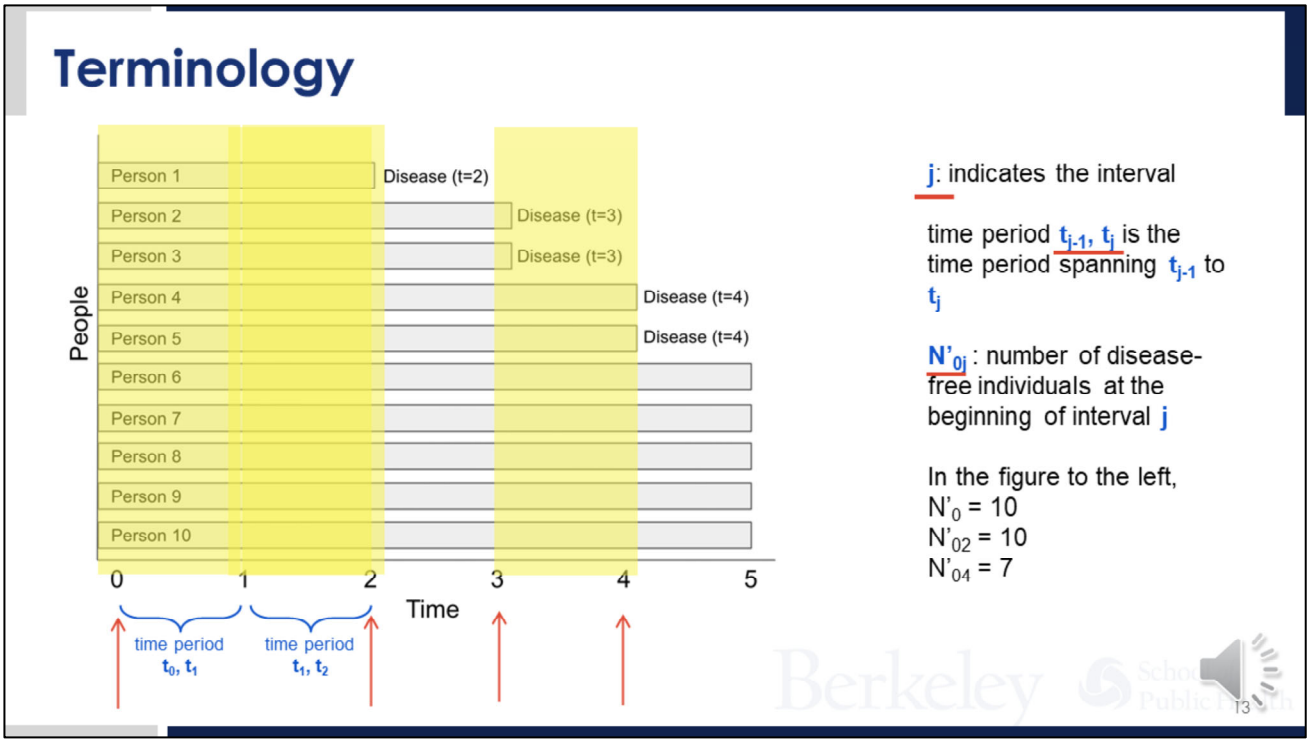
*So again, it's appropriate for incomplete follow up, and *here's an intuitive breakdown of what the actuarial method involves. First, we're going to break up our follow-up period into smaller time intervals. And within those intervals that are shorter periods of time, *we're going to treat our population as a closed cohort.

So let's say our study is a study of Californians, and there's tens of thousands of participants. It's an open population. People can move into California. They're born in California. People can exit California, leave California. And let's say we have 10 years of data on these Californians. Well, that's a pretty long period of time. It certainly would not be reasonable to assume that we had complete data on people over a 10-year period or even that there weren't large changes in disease trends over that 10-year period. So what we need to do, intuitively, is to break up that 10-year period into much smaller time intervals, in which it's safer to assume that we're actually dealing with a closed cohort. Or you could think of it as a closed population within that cohort.

*And then within those smaller intervals, you'll estimate the risk in each interval, making an assumption that, when people withdrew from the study, it occurred halfway through the interval. And that's because, while it won't be true for most people individually, it's a reasonable assumption that the average withdrawal will haven halfway though a period. And for this reason, because we're making this assumption, *we want the interval length to be relatively short, and *we can think of the risk for each interval as a conditional probability. It's a probability that conditions on whether a

person was at risk. In other words, they were alive and not censored at the time of the disease event.

So after we're done with step three, *we're going to have estimates of risk for each interval, and then we want to put them together to get our cumulative incidence over the whole period. So let's say we estimated monthly risk of disease for our 10-year long study, so we would have 120. Right? Because we have 10 years and 12 months per year, so we'd have 120 estimates of risk in small, small time intervals. And then we would want to put those together to get our cumulative incidence over the 10-year period. So this is, in a big picture sense, the process that I'm about to walk you through.



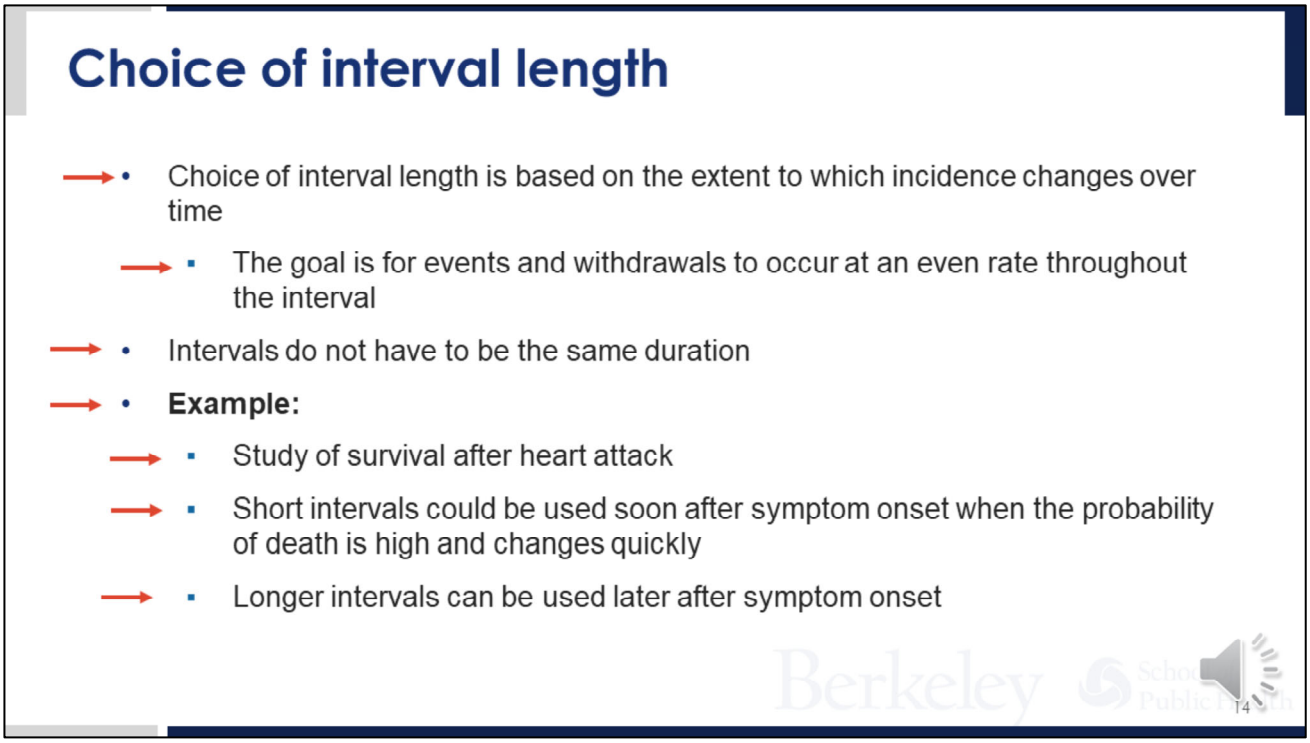
Let's go over some terminology. So we're going to use the *index j to indicate which interval we're talking about, and then we can use the *letter t to refer to our time period of interest. So t sub j is a specific time interval, and t sub j minus 1 is the time interval before t sub j.

**So if we look at our graph, here on the bottom left, I've indicated how we would notate time period t sub 0 comma t sub 1. So this is just this time period spanning time 0 to 1. *And then from the time period 1 to 2, we would write this as t sub 1, t sub 2. All right? So that's our t and our j notation.

*Then, we also have this modified version of the notation we saw earlier, so we have N prime sub 0 j. This means the number of disease-free people at the beginning of interval j. So for simple cumulative incidence, it was actually at the beginning of the study. But in this actuarial method you're going to learn that it's all specific to the interval when we calculate things, and so that j is going to indicate that it's within a specific interval. *So for example, if we're looking at the interval from 1 to 2, we would want to know the disease-free individuals at the beginning of interval 1 to 2, which would be the number of people without disease at time 1.
So let's calculate a few of these. *So in the figure to the left here, N prime 0-- so at the very beginning of the study-- there's 10 people who are at risk of disease. So it's equal to 10.

*And then at time 2, so N prime 0 2, that's the number of disease-free individuals at

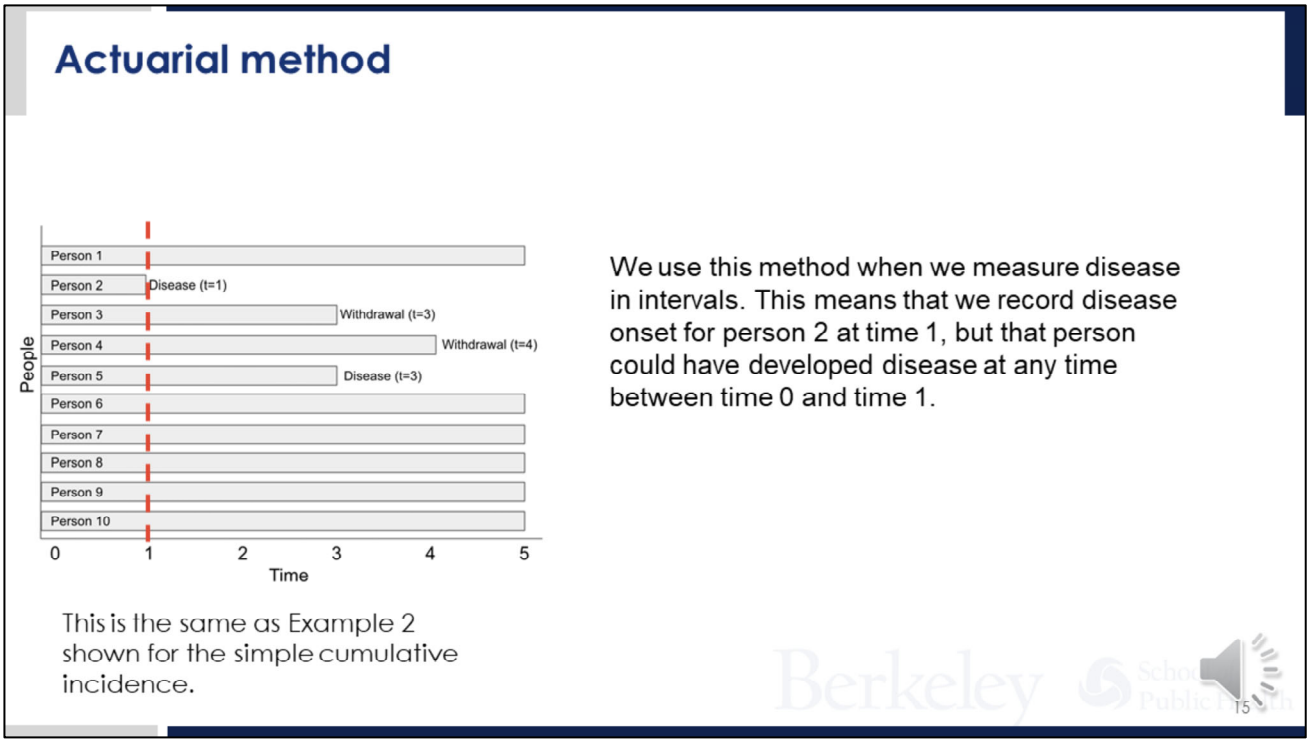
the beginning of interval 2. The number is also 10, because as we can see here, there were 10 people who were at risk. *And then if we look at N prime sub 0 4, right here, *what we can see is that persons 1 through 3 develop disease before time 4, and persons 4 and 5 developed the disease within the period 4. But we're not going to exclude them from the denominator, because the disease occurred during that time period. So if we're thinking about our time period from 3 to 4, we're going to say that, at the beginning of that interval, there were 7 people at risk of disease.



How do we decide how long the interval should be? Well, this is based on the extent to which we think the incidence changes over time. So we want to define our intervals in a way that makes the events, AKA the disease onset, and withdrawals occur at an even rate throughout the interval. So we wouldn't want the interval to be so long that all the disease events were occurring at the beginning of the interval, and all the withdrawals were happening at the end of the interval.

So a year, for example, might be way too long for most diseases. It would make it very hard for us to have a good balance in the number of events and withdrawals throughout that year-long period. So we'd probably want to pick a shorter interval duration, but if you have a very long study, you actually don't need to use the same duration of interval throughout the entire thing.

So here's an example, if we're looking at a study of survival after a heart attack. So we could use very short intervals for the period of time after symptom onset, when the probability of death is much higher, and the probability is changing very rapidly. But then, after more time has passed since symptom onset, we could have a longer interval, and that would be appropriate for this particular study.



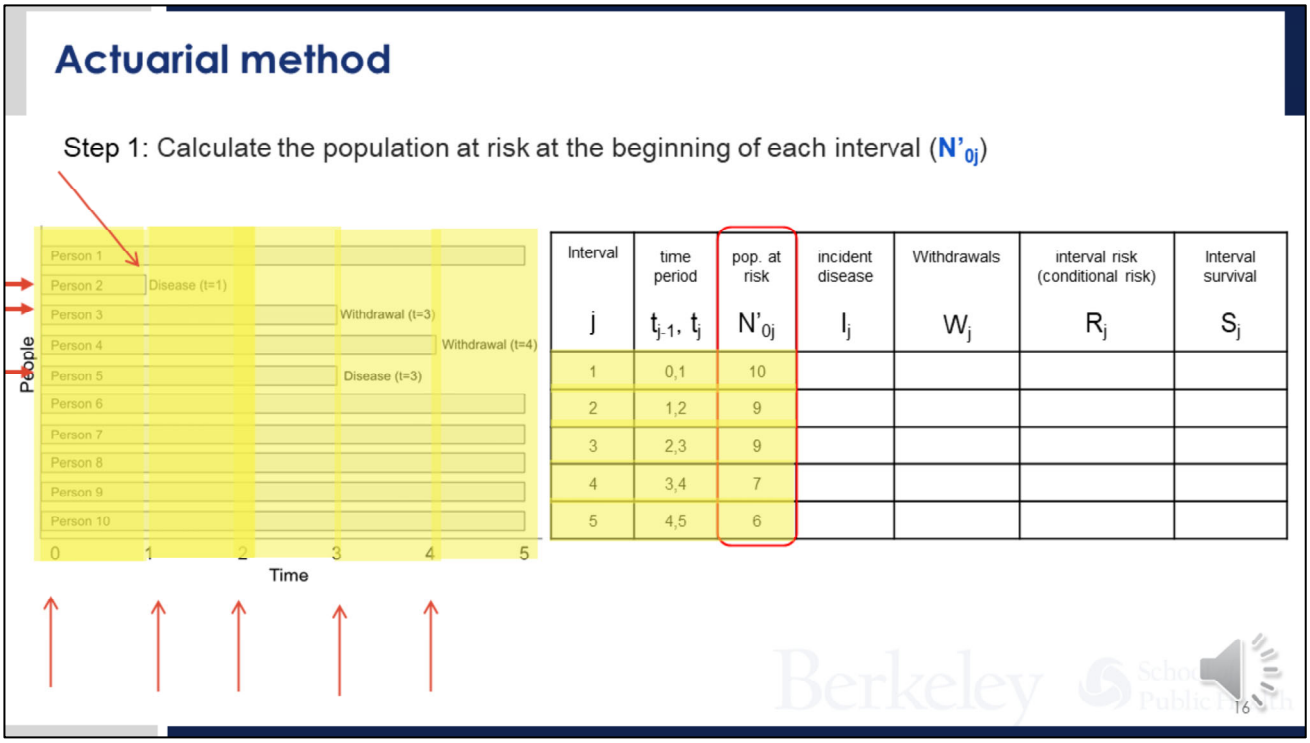
I want to make a brief note about the way this picture is drawn. So this is the same example that we used for a simple cumulative incidence, when we said don't use the simple cumulative method, because there's withdrawal. So we're going to perform the actuarial method now on that same example, but I want to just point out the way we've drawn this picture.

We see that the disease and withdrawal occurrences are actually happening right at the specific tick marks for time. Right? So for person 2, their disease onset occurs right at time 1. We never see anybody have a disease in between time 1 and time 2. And this is purposefully drawn this way, because in some studies, we actually aren't able to continuously measure disease status.

So what's an example of that? Well, if you're doing a study in a hospital, where a person is hooked up to monitors 24/7, you're able to continuously monitor their status, if disease is defined based on whatever the monitor is reading. But in most studies, that's actually not the case. So another example might be a study where patients have to go in once a week to have some bloodwork done or have some other test performed to assess whether they have developed the disease. And we actually wouldn't know what's happening in between those weekly visits.

So if a person comes the first week of the month and doesn't have the disease, and they come the second week of the month, and they do have the disease, it actually doesn't mean that the disease onset was right at the second week of the month. It

could have actually been any time between the first and second week. So these pictures, just a caveat for you to keep in mind, is that this is sort of artificial. There could be disease events or withdrawal events occurring in between these number ticks here, and that's something that you really need to remember and understand as we think about the assumptions behind the actuarial method.



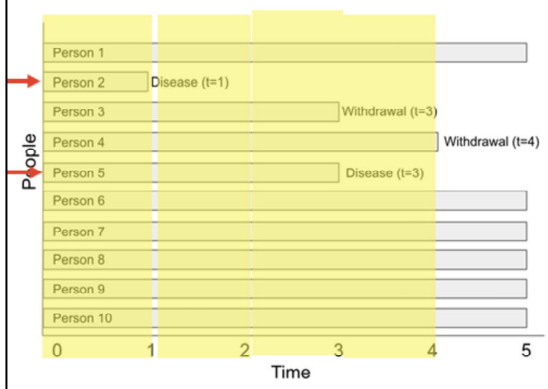
The first step of the actuarial method is to calculate the population at risk at the beginning of each interval. So let's say that we're just going to break up this time of follow up from 0 to 5 into one time unit increments. *So interval j equals 1 spans the time period 0 to 1. *Interval j equals 2 spans the time period 1 to 2, and so on.

*So then for each of these rows, we need to calculate the population at risk at the beginning of that interval. *To get the population at risk for the first interval, from 0 to 1, *we look at time 0, there were 10 people without disease. *So the answer is 10. *And then when we look from 1 to 2, we see that a person had developed the disease at time 1. So they were no longer at risk of developing the disease between time 1 and time 2, but everybody else was still at risk. *So the population at risk is 9. *And nothing changes from 2 to 3, so the population at risk is still 9.

*And then, when we look from time 3 to time 4, *that person who is lost to follow up, person 3,* and the persons 2 and 5 who develop the disease prior to or at the beginning of time 3 are not at risk anymore between time 3 and time 4. So we have 7 people who are at risk of disease. *And then from time 4 to 5, it drops to 6, because we had one more person who was lost to follow up or who withdrew at time 4, and that is person 4. All right, so we've now filled in our population-at-risk column.

Actuarial method

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



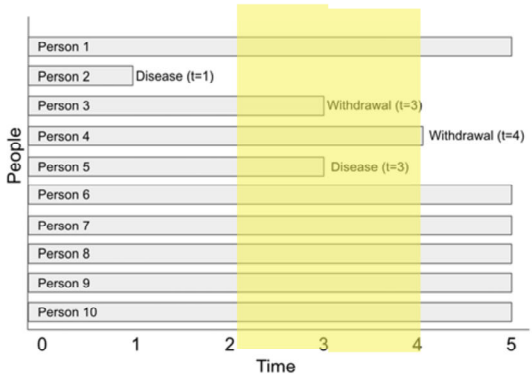
Interval j	time period t_{j-1}, t_j	pop. at risk N'_{0j}	incident disease I_j	Withdrawals W_j	interval risk (conditional risk) R_j	Interval survival S_j
1	0,1	10	1			
2	1,2	9	0			
3	2,3	9	1			
4	3,4	7	0			
5	4,5	6	0			

Next, we need to count the number of incident cases within each interval. So this is pretty straightforward. We see that in interval 1, there was 1 person who developed the disease between 0 and 1.

Now, again, because of the nature of measurement in this hypothetical study, remember this person 2 who developed disease at time 1 actually could have developed it before time 1. But that's when we measured it, so we're going to assign it to interval 1. Between time 1 and 2, nobody developed disease, so that's 0. Between time 2 and time 3, we have one new case of disease in person 5, and then no more subsequent cases. And so we end up with only two rows with an incident disease.

Actuarial method

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



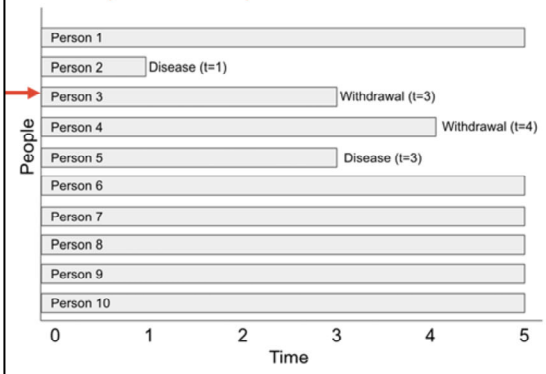
Interval j	time period t_{j-1}, t_j	pop. at risk N'_{0j}	incident disease I_j	Withdrawals W_j	interval risk (conditional risk) R_j	Interval survival S_j
1	0,1	10	1	0		
2	1,2	9	0	0		
3	2,3	9	1	1		
4	3,4	7	0	1		
5	4,5	6	0	0		

Step three is to count the number of withdrawals within each interval. So it's very similar to what we just did for incident disease. We didn't see any withdrawals until interval 3, *and from time 2 to 3, we have one withdrawal occurring. *And then from time 3 to 4, we have one withdrawal occurring, and so we fill in the table accordingly.

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	

Now, we're going to calculate the risk within each interval, and again, this is an estimate of the risk. This isn't necessarily the real risk, but we're calling it risk, because within these intervals, we're treating our steady population as a closed population. So that's an assumption that we're making in this method.

And another assumption we're going to make is that withdrawals are occurring evenly throughout each interval. So what do we mean by that? *Well, when we look at this formula for risk here, *what we see is we have incident cases in the numerator. That's the exact same as in the simple cumulative method. Then, in the denominator, we have the *population at risk at the beginning of the interval.

So so far, this is the exact same as a simple cumulative method, but what's the difference? *This minus Wj over 2. This part right here is accounting for the withdrawals that occurred in each interval. And the reason we divide is by 2 is that we want to make the assumption that the withdrawals are occurring evenly in the interval. So the probability of the withdrawal occurring is equally likely at the beginning and end of the interval.

*Let's look at person 3 in this figure. So person 3 is indicated as a withdrawal at time t equals 3, but we actually don't know the true time of withdrawal. It really could have at anytime between 2 and 3. So what we're going to do is assume that that withdrawal could have happened halfway through the interval. That's why we're dividing that withdrawal by 2.

*So let's calculate it for row 1. So we have 1, which is our incident disease, divided by 10. We're not subtracting anything, because there are no withdrawals, so that's equal to 0.1.

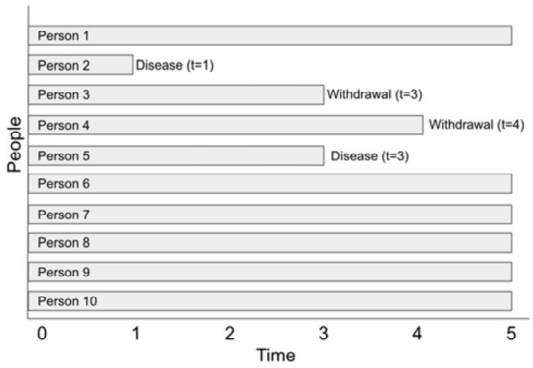
*Now, I'd like noted here that you should keep three decimal points. Why is that? Well we're going to be working with a bunch of decimals here, and in reality, it'd be better not to round it all. But just to make our lives easier, for the purpose of learning this, we're going to round to three decimal points when we calculate the interval risk and the interval survival.

And then, by keeping three decimal points, we're less likely to have a rounding error at the very last step. So in these videos, in your homework, on the exams, please, please, please, always keep three decimal points until the very end. Don't round before then.

*All right, let's look at interval 3. So let's calculate the interval risk for that row. So we have 1 incident case divided by 9 is our population at risk, and then we had 1 withdrawal, and we're going to divide that by 2. So it's 1 over 9 minus 1 over 2, which is 0.118. Same approach for the row below that. So that's our interval risk.

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	

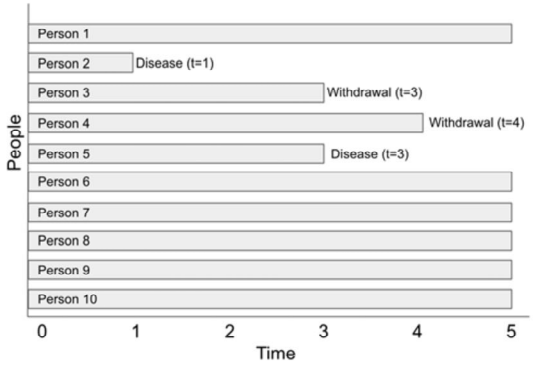
And now here is our interval survival.

So what does this mean? Well, when we think about the risk of developing the disease, it's really the opposite concept of the risk of surviving. So we can say that, if you survived, it means you didn't get the disease. So S sub j is the interval survival, and that's equal to 1 minus the interval risk of R sub j, because R sub j is a probability that spans 0 to 1. *To calculate the interval survival, we just subtract 1 minus the interval risk, and that's pretty straightforward.

Actuarial method

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

Keep 3 decimal points!!



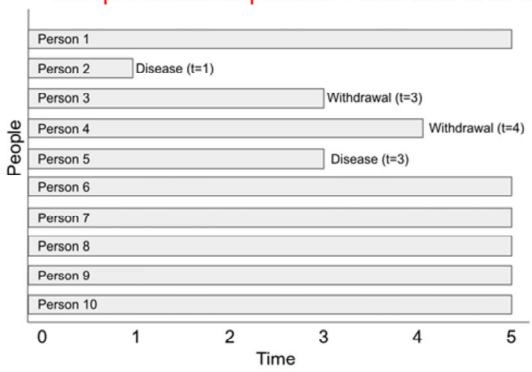
Interval j	time period t_{j-1}, t_j	pop. at risk N'_{0j}	incident disease I_j	Withdrawals W_j	interval risk (conditional risk) R_j	Interval survival S_j
1	0,1	10	1	0	1 / 10 = 0.100	0.900
2	1,2	9	0	0	0 / 9 = 0.000	1.000
3	2,3	9	1	1	1 / (9 - 1/2) = 0.118	0.882
4	3,4	7	0	1	0 / (7 - 1/2) = 0.000	1.000
5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

And it's shown in this final column, right here.

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)

22

Step six, this is the final step. We're going to calculate the cumulative incidence. So right now, we have interval survival, and we have interval risk. We want the cumulative incidence across the whole follow-up period. *We're going to use something called the product limit formula, and we'll come back to how this is derived in a minute. It looks a little bit complicated, but it's really not too bad.

*So here CI is cumulative incidence. Sub t0 tj, this means we're interested in the period from time 0 to interval j. And generally speaking, when we write it this way, we mean the whole follow-up period, *and that's equal to 1 minus the grand product of 1 minus R sub j which is the interval risk.

What is the grand product? *It just means take this part here, after the product sign, and for every single interval j-- right, so from 1, 2, 3, 4, and 5-- we're going to multiply all those together. *And then we can simplify that to just 1 minus the grand product of the survival.

*This calculation is shown at the bottom of this slide. We see 1 minus the product of 0.9, which is shown in the first interval survival cell, times 1, which is shown in the cell in the second row of the table, times 0.882 times 1 times 1, and that equals 0.2062, which we can round down to 0.206. So that's our result. That's the cumulative incidence for this whole follow-up period using the actuarial method.

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

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

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

$$S_j = 1 - R_j$$

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

23

So I just went over this. I'm not going to talk through it, but I'm putting this slide here in the video for your reference. So that when you come back to this later, you can see all the steps in one place.

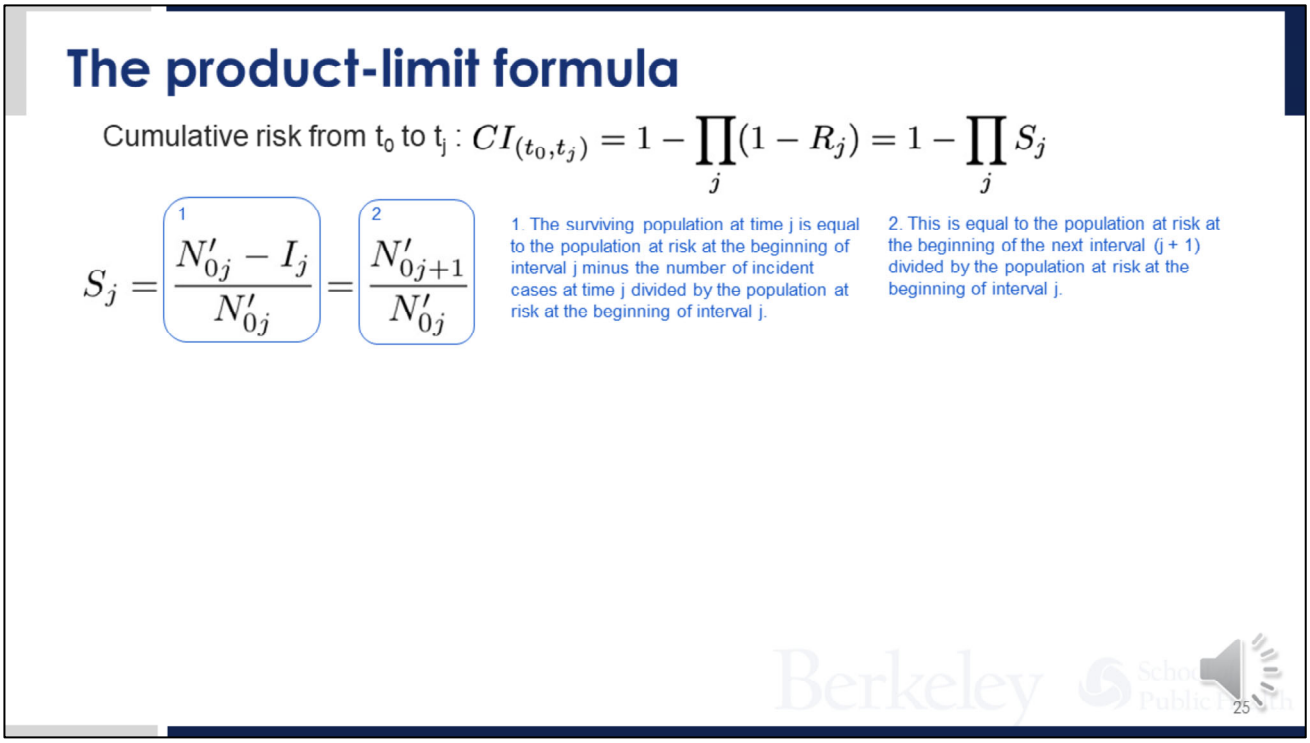
The product-limit formula

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

Now, I'd like to talk about that product limit formula. It seems a little bit out of thin air, but it's actually relatively straightforward, even though the math does look a little bit intimidating. I'd like to just talk through where it comes from.

We don't expect you to know this. We do expect you to know the actuarial method and how to calculate it. You don't need to be able to derive the product limit formula, but I'm just showing this to you for your own information.

So here's that same formula I showed you in the previous few slides.

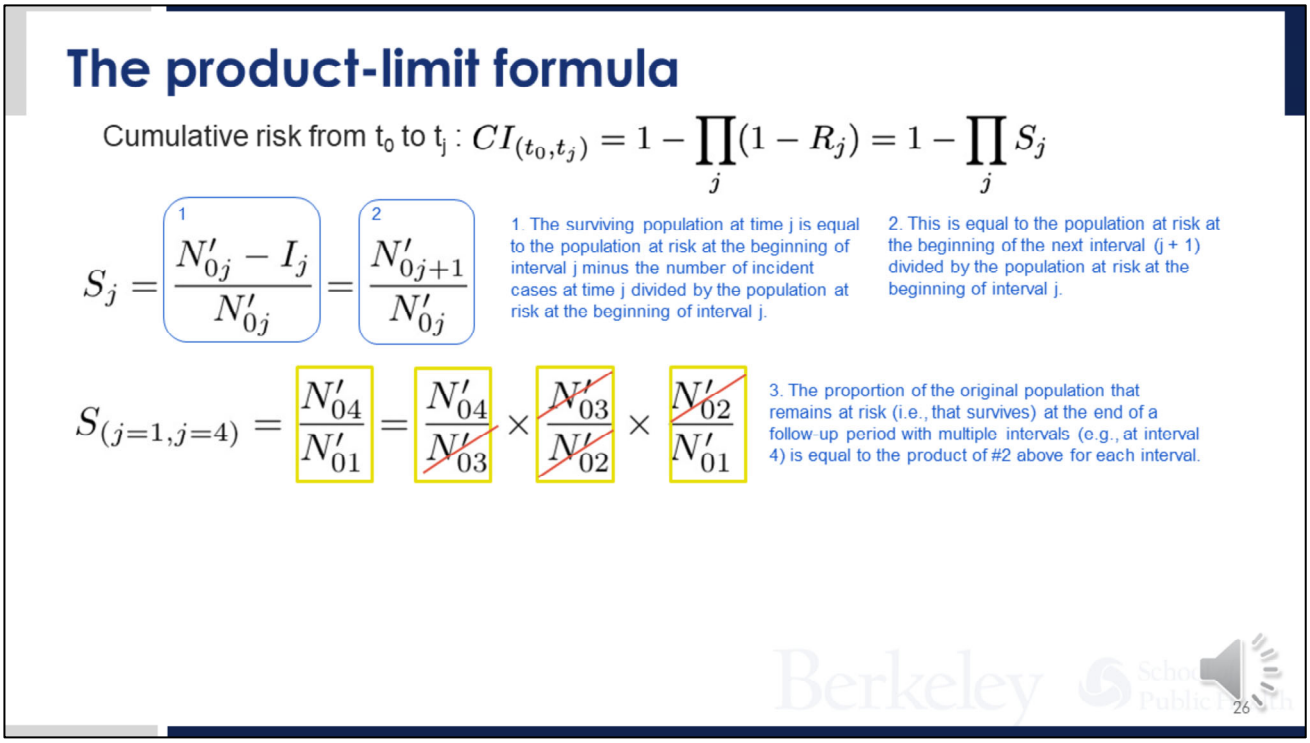


And let's start by thinking about what survival in an interval is actually equal to. So in this box number one, we have the surviving population at time j is equal to the population at risk at the beginning of that interval j. Right? So that's N prime sub 0j, minus the number of incident cases at time j, which as I sub j, divided by the population at risk at the beginning of interval j, which is in the denominator.

What does this mean? We're just saying that the people who survived at interval j is equal to the proportion of people who didn't get the disease. We are looking at the people who didn't have the disease to begin with in that interval, and we're just subtracting out the incident cases in the numerator. So that's the survival.

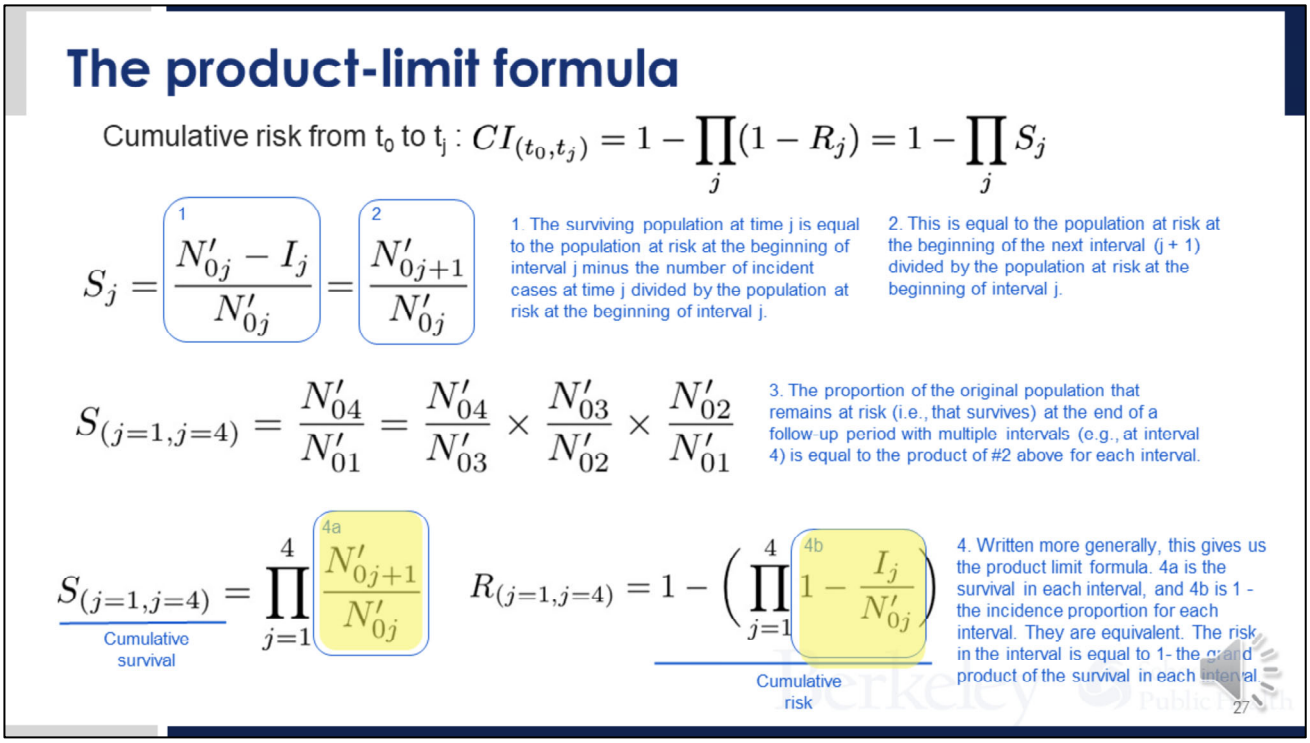
Now, in step two, we're doing something fancy with the notation. We're basically saying that that quantity in step one is equal to the population at risk at the beginning of the next interval divided by the population at risk at the beginning of interval j. So nothing has changed in the denominator.

But we're just saying that in the numerator, we're changing up the notation in the subscript, because the end of interval 1 is the same as the beginning of interval 2. So that's really all this is saying, and this is just a convenience from a math standpoint. That's why we're doing that.



Step three, we can break up the survival, and let's say, we're only looking at 4 intervals, from j equals 1 to j equals 4. So moving from step two, *we're looking at N prime sub 0 4 divided by N prime sub 0 1. That's the population at risk at the beginning of interval 4 divided by the population at risk at the beginning of interval 1.

That can be factorized as follows. It's the product of the *proportion of people at the beginning of interval 4 divided by the portion of people at risk at the beginning interval 3 times *the proportion of people at risk at the beginning of interval 3 divided by the proportion at risk at the beginning of interval 2 and* so on. So really this is looking quite complicated, but it's really a mathematical concept from arithmetic. *We can cancel N prime 0 3 and N prime 0 3 and N prime 0 2 and N prime 0 2. We're just basically expanding this out, because it's going to link us back to this concept of the grand product over the interval.



So again, we're looking at the survival proportion from interval 1 to 4, j equals 1 to j equals 4. We can rewrite this previous step as a grand product from j equals 1 to 4, and see here, in box 4a, it looks a lot like the box on number 2 above. So this is basically what's giving us the product limit formula. So 4a is the survival in each interval multiplied together. 4b is 1 minus the incidence proportion in each interval, and these are basically equivalent.

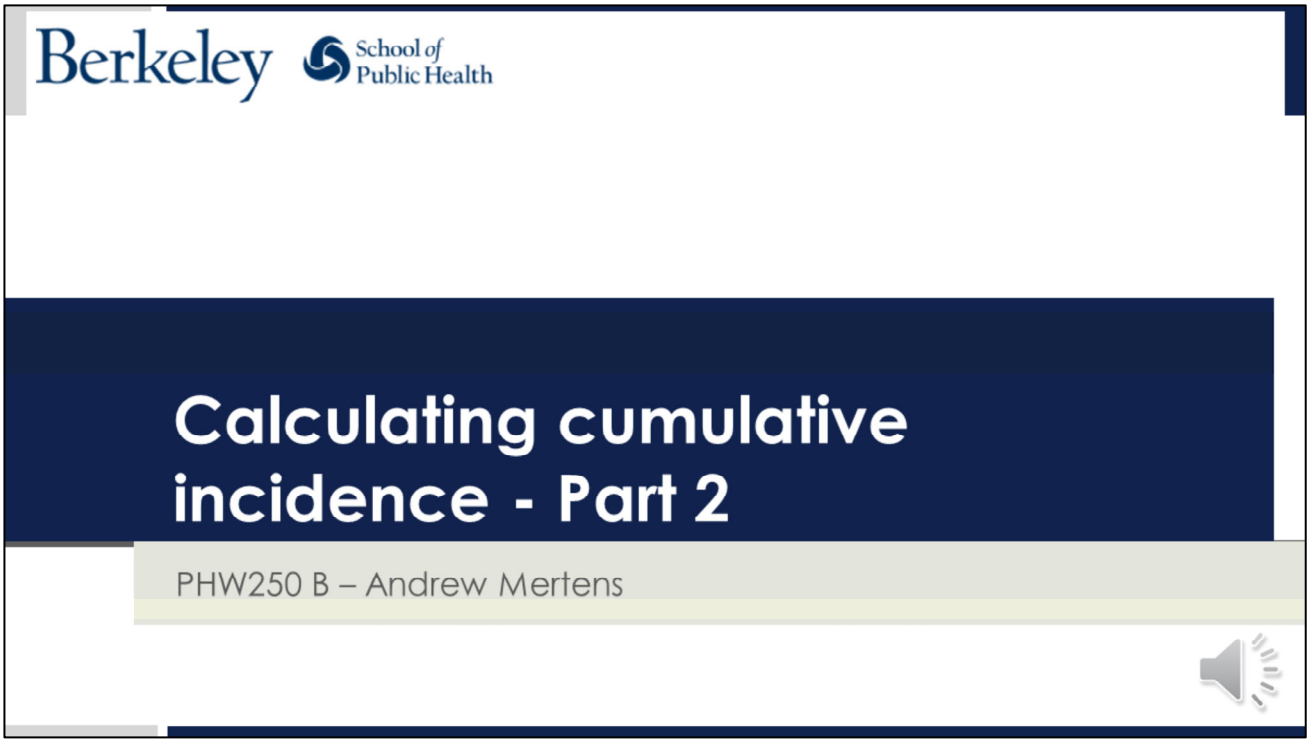
So what we've shown is that the risk's equal to 1 minus the grand product of the survival in each interval. And then, if you look at the top of the slide, there's just a little bit of simplification with the 1 minus part that gives us the formula that we used. So you don't need to know this, but we just wanted to show you where this comes from. It's not something that you need to be able to derive on an exam.

The next video will cover the remaining topics in our outline

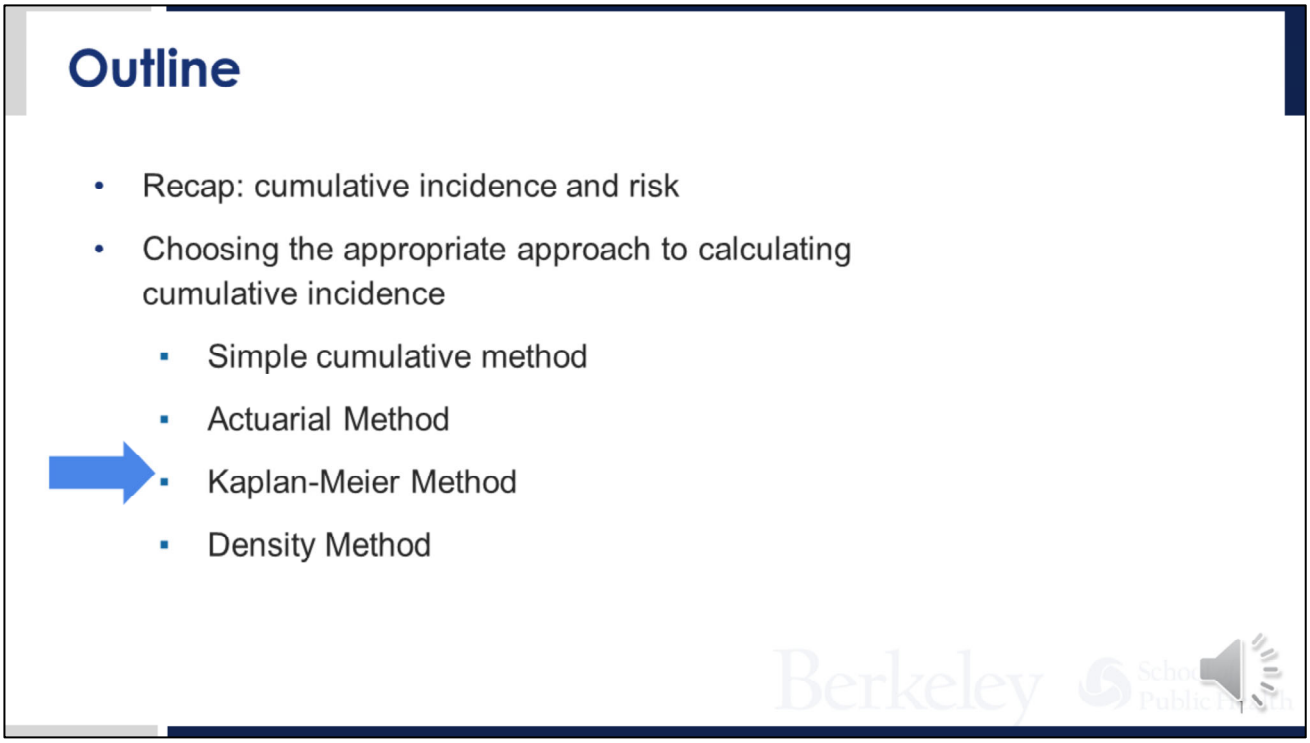
- ✓ Recap: cumulative incidence and risk
- ✓ Choosing the appropriate approach to calculating cumulative incidence
  - ✓ Simple cumulative method
  - ✓ Actuarial Method
  - Kaplan-Meier Method
  - Density Method

28

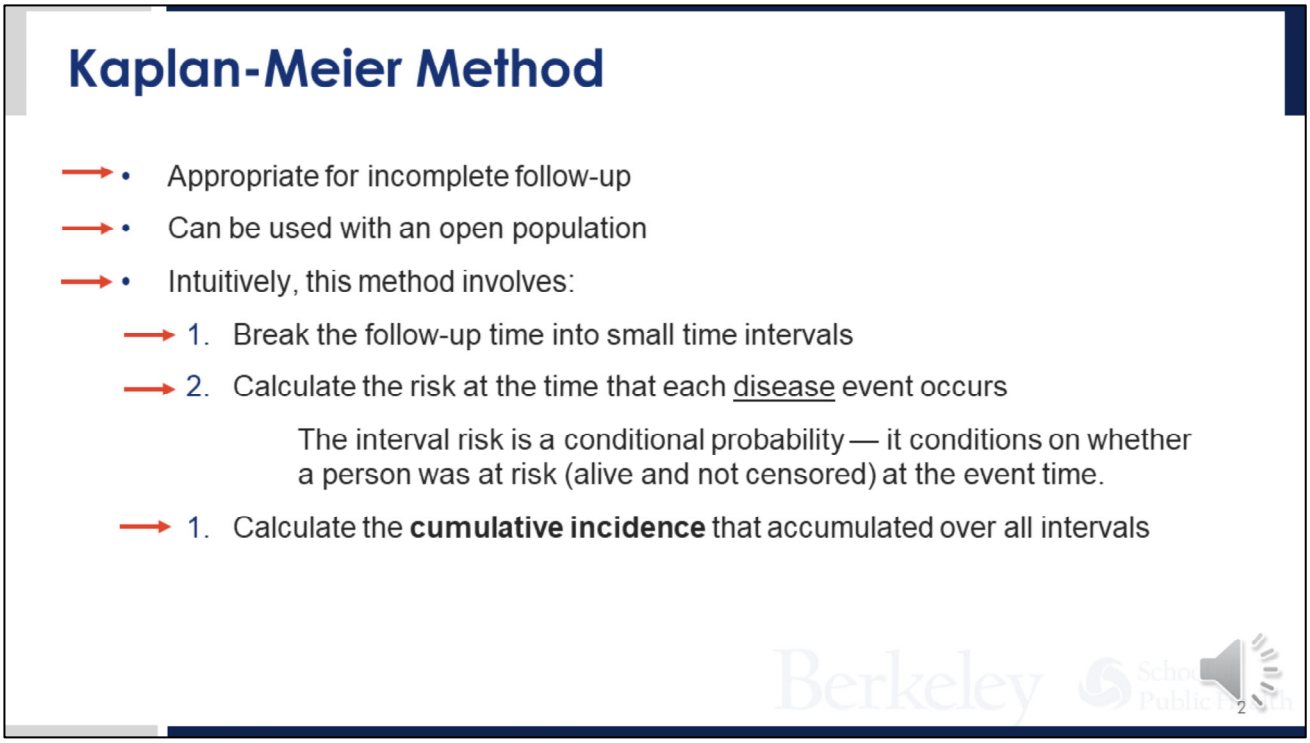
OK. In this video, we've covered cumulative incidence and how it relates to risk. We've covered the simple cumulative method and the actuarial method, and in the next video, we'll finish up with Kaplan-Meier and the density method.



This is the second of a two-video series on calculating cumulative incidence.



In the last video, we talked about the difference between cumulative incidence and risk. And we covered how to choose between different methods of calculating cumulative incidence-- the simple cumulative method and the actuarial method. And in this video, we'll start with the Kaplan-Meier method and close with the density method.



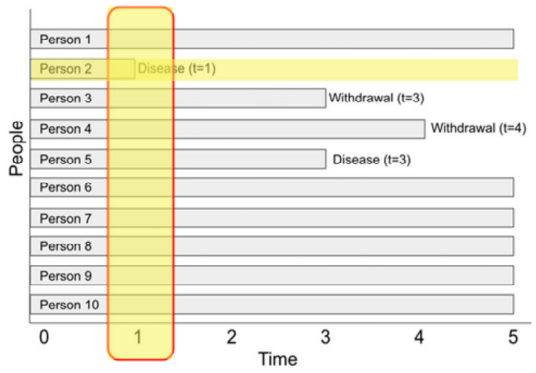
The Kaplan-Meier method is also appropriate when your population has incomplete follow-up, meaning some people were lost to follow up, there was attrition, or there was censoring in your steady population. And it's a method that can be used with open populations.

Here's a intuitive breakdown of what the method involves. In the first step, we'll break our follow-up time into smaller time intervals just like we did in the actuarial method. In the second step, we'll calculate the risk at the time that each disease event occurs. So the timing of the disease event is really what's going to determine our interval breakdown. This is the key difference between the actuarial method and the Kaplan-Meier method.

And again, we can think of the risk within each disease event interval as a conditional probability that conditions on whether a person was at risk. In other words, they were alive and not censored at the time of the disease occurrence. So after we get a bunch of interval-specific risk estimates, *we'll calculate the cumulative incidence across all these different 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.



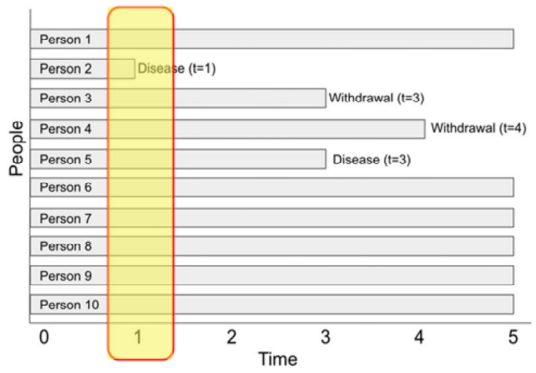
Time j	pop. at risk N'_j	incident disease I_j	interval risk (conditional risk) R_j	interval survival (conditional surv.) S_j
1	10			

In the first step, we need to identify the disease events and calculate the population at risk at the time of each event. And that's indicated by N prime sub j. And the person who developed the disease at that time is included in N prime sub j. *So the first disease event was in person two at time one. And that will define our first interval.

So again, this is quite different from the actuarial method. We're filling in our table and our j column based on the timing of disease events. *So for the first time, time j where we have person two developing disease, there are 10 people at risk of disease, because we, again, we're including person two in that denominator.

Kaplan-Meier method

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

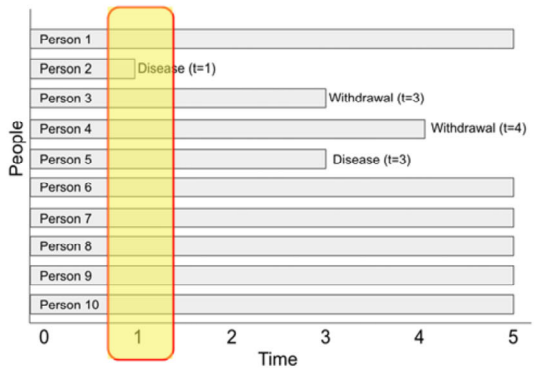


Time j	pop. at risk N'_j	incident disease I_j	interval risk (conditional risk) R_j	interval survival (conditional surv.) S_j
1	10	1		

Then staying within the same time interval, we're going to count the number of incident cases at time j. So that's I sub j. And there was only one person who developed the disease at that time.

Kaplan-Meier method

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



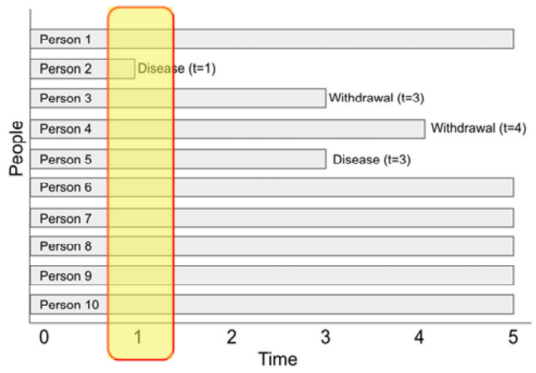
Time j	pop. at risk N'_j	incident disease I_j	interval risk (conditional risk) R_j	interval survival (conditional surv.) S_j
1	10	1	1/10 = 0.100	

Then we're going to calculate the risk at that time of developing the disease. And again, we're going to keep three decimal points, just like we did in the actuarial method.

So that's 1 over 10 equals 0.100.

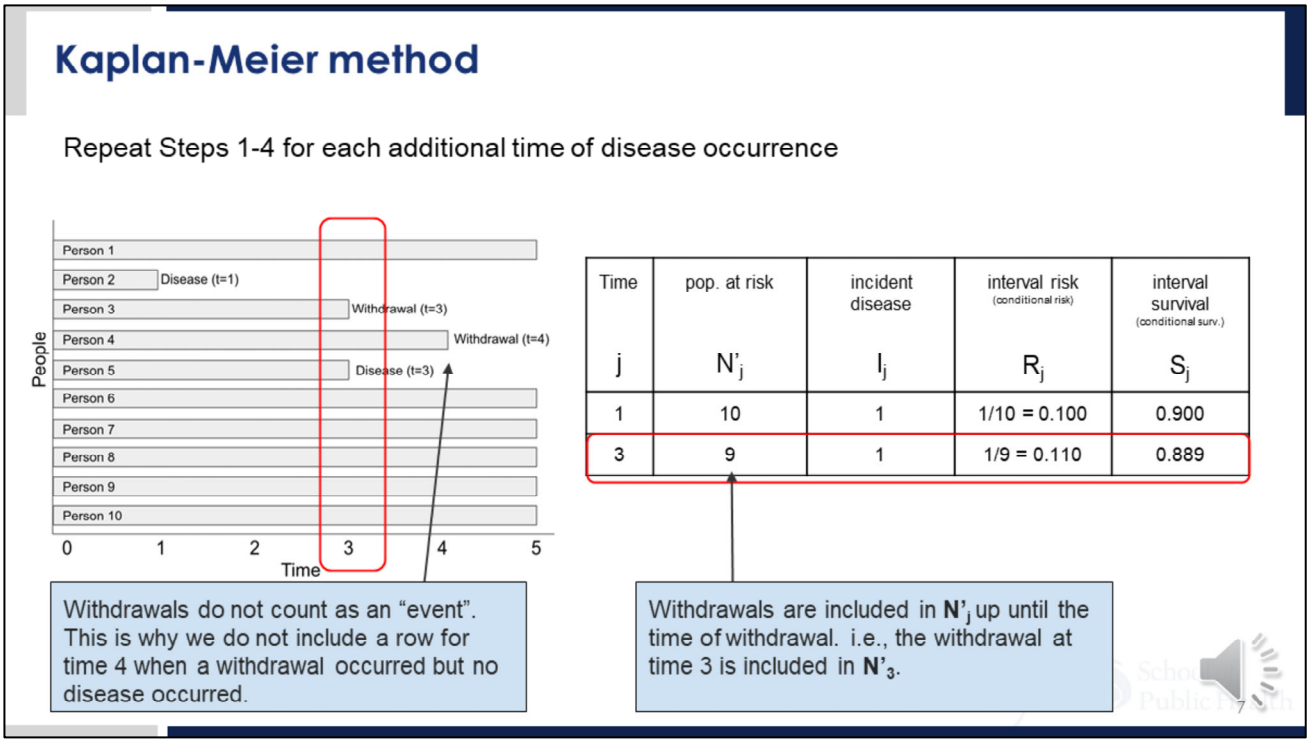
Kaplan-Meier method

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

And then we'll calculate the survival at that time. So that's 1 minus the risk, which is 0.900. So we've now done this for the first disease event. Then we're going to repeat those steps for each additional time of disease occurrence. So in this small example, there was only one other disease occurrence in person five at time three. So our table's only going to have two rows-- one for the disease occurrence at time one and one for the disease occurrence at time three. And I just want to make a couple of notes here.



At time three, we see that there was also a withdrawal. So person three was withdrawn from the study at that time. We don't consider withdrawals to be an event. So when we say event, we just mean disease events. And that's why the table only has rows at times one and three and doesn't include a row for time four when there was a withdrawal but no disease.

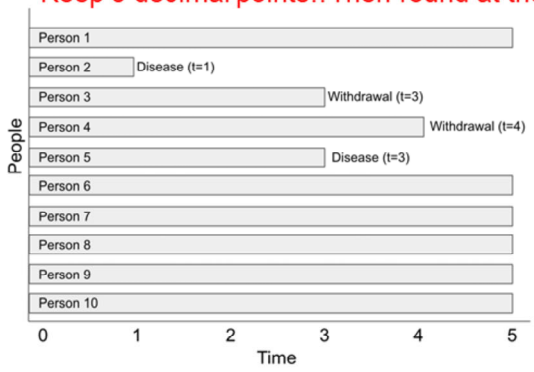
And then I also just wanted to know at time three-- again, when we have one disease event and one withdrawal, they are included in N prime sub j up until the time of the withdrawal. So in other words, at time three, we do include the person who withdrew at time three in N.

The only person who is excluded is the person who developed the disease at the prior time period. So person two doesn't contribute to the population at risk at time three. And so in row two of the table, the calculations for interval risk and survival have been completed keeping these rules in mind.

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

Just like in the actuarial method, the last step of the Kaplan-Meier method is to *apply the product-limit formula, because we now have these interval risks and interval survivals. And we need to get the cumulative incidence across all the intervals. So here, we have that same formula. *And it's going to be the cumulative incidence from time 0 to time 5 equals 1 minus the grand product of 0.9 times 0.889. *There's only two numbers to multiply here. That gives us 0.1999. And we can round that to 0.200.

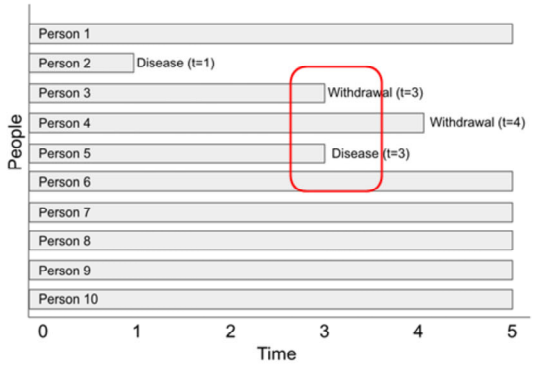
Kaplan-Meier method steps

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

So I have, again, put all the steps here. I'm not going to go over them since I just did so. But here they are so that when you're reviewing these slides, they're all in one place.

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

Now, let's compare the results that we got from the actuarial and the Kaplan-Meier methods. And what we can see is that they're slightly different. In the actuarial method, we had an estimate of 0.206 and the Kaplan-Meier method, an estimate of 0.200.

And this is because the actuarial and Kaplan-Meier methods differ when, at least, one interval has both a withdrawal and a disease occurrence. And so that happened in this example at time three. Let's dig a little deeper to see why.

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

So this is a very busy slide. *But the table on the left is the actuarial method. *The table on the right is the Kaplan-Meier method.

And there's two key differences to notice. First, in the actuarial method, we're looking at all of the intervals. We have five rows. But in the Kaplan-Meier method, we're only looking at times when a disease occurs. We only have two rows. That's one key difference.

The second key difference is related to how withdrawals are handled. For time three, remember there was that withdrawal at the same time with the disease event. *So the actuarial method is basically going to assume that that withdrawal occurred halfway through the interval. And that's why 1 is divided by 2 in this particular cell here in the denominator.

*Now, if we look at the Kaplan-Meier method on the right, we see that it's just 1 over 9. These two methods are handling the withdrawals a little bit differently. Again, the actuarial method is assuming that the withdrawal occurred halfway through the interval.

The risk is 1 over 9 minus 0.5, which is 1 over 8.5. But in the Kaplan-Meier method for the same row, the risk is 1 over 9. And so it's not assuming that the withdrawal occurred halfway through. It's letting that person sort of be followed up all the way through that particular interval.

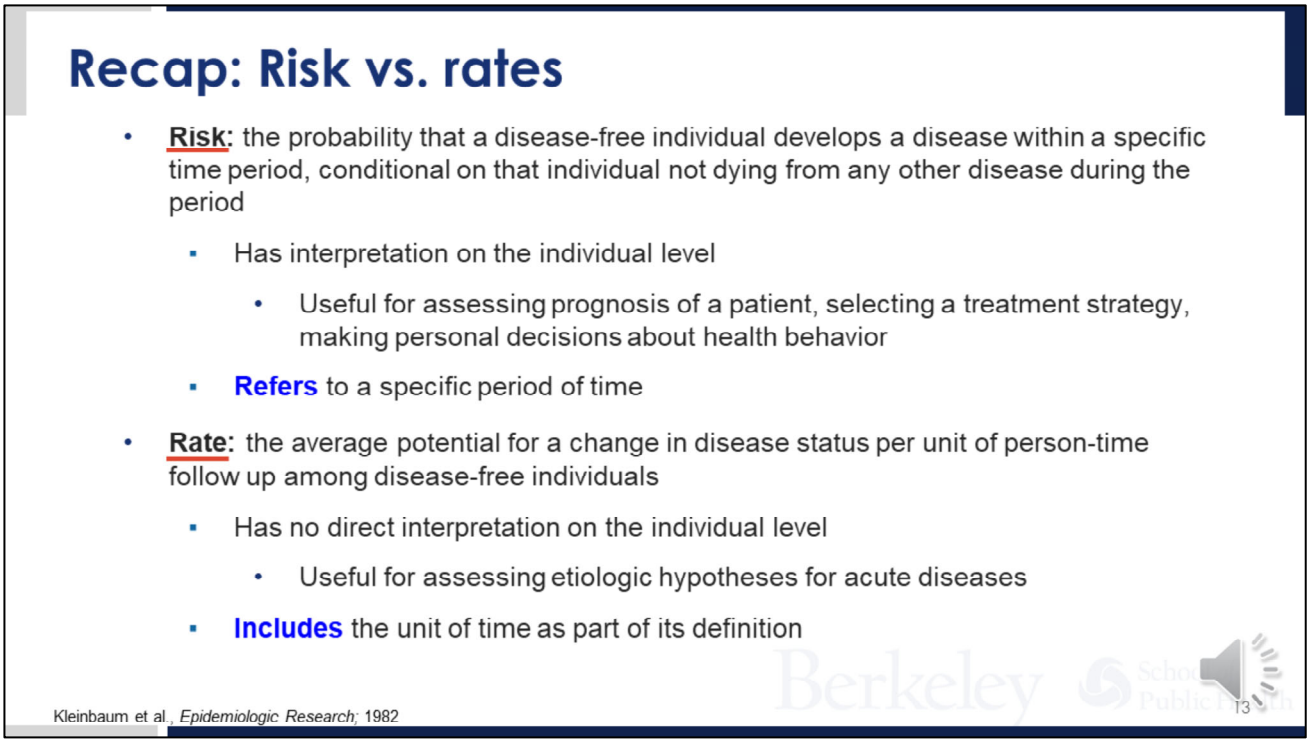
So the risk is a little bit higher for that interval in the actuarial method than it is in the Kaplan-Meier method. So generally speaking, these methods agree quite well. But again, if you have a withdrawal and a disease occurring at the same time, you will have some divergence in the results for these two methods.

Outline

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

12

Now, let's move on to the density method.



Again, this is a method that we use when we're interested in converting rate to risk. So I'd like to briefly recap the difference between risk and rates. Risk is the probability that a disease-free person or individual develops a disease in a specific time period conditional on that person not dying from any other disease during the period.

It's interpreted at the individual levels. This is very key. And it's useful for making prognosis of a patient in a clinical setting, for example, and selecting a treatment strategy, making a personal decision about changing health behavior. And its referenced to a specific point in time. Rate is the average potential for a change in disease status per unit of person time follow up among disease-free individuals.

So risk is referring to a specific period of time, but the units do not have time, whereas a rate includes time as part of the measure. And it has no interpretation at the individual level. But it is useful for assessing the rate of disease onset. And so it's helpful for etiologic hypotheses, especially for acute diseases with short durations.

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

14

So we can use the density method if we have rates, and we want to convert them to cumulative incidence. And we can use this method when we have incomplete follow up and we have an open population. The formula that we're going to show you, though, does assume that the disease is rare. So this is something to keep in mind.

Here is an intuitive breakdown of what the steps entail. Just like in the last two methods, we're going to break up the follow-up period into smaller time intervals. We'll calculate the incidence density within each interval. We will then convert the incidence density to an estimate of risk within each interval using a formula that links risks and rates. And then we'll calculate the cumulative incidence over all those intervals.

Formula to link cumulative incidence and incidence density

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

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

Rothman ME3 Page 43

First, I want to show you this formula that links cumulative incidence to the incidence density or the rate. So if R stands for risk and ID stands for incidence density, here, what we have is that the risk in an interval j is equal to the incidence density in that interval times delta t. So that's the duration of follow up time.

So if we multiply this through incident cases over person time at risk--that's our definition of incidence density-- multiply that times follow-up time. And we cancel out the time units. And what we get back is incident cases over persons at risk. So what we see at the top here, the first line is a rate, because our denominator is person time multiplied by time. And then a second line is a risk because it's a probability. We're dividing people who develop the disease by people who are at risk. And it's unitless.

Formula to link cumulative incidence and incidence density

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

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

$S_j = 1 - ID_j \triangle t_j$

Now calculate S, using S = 1 - R

And as we've discussed, we can calculate the survival, which is 1 minus the risk. And so survival at interval j is equal to 1 minus the incidence density interval j times delta t sub j from the last step. So we've just plugged in R sub j from the first step into R the second step.

Formula to link cumulative incidence and incidence density

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

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

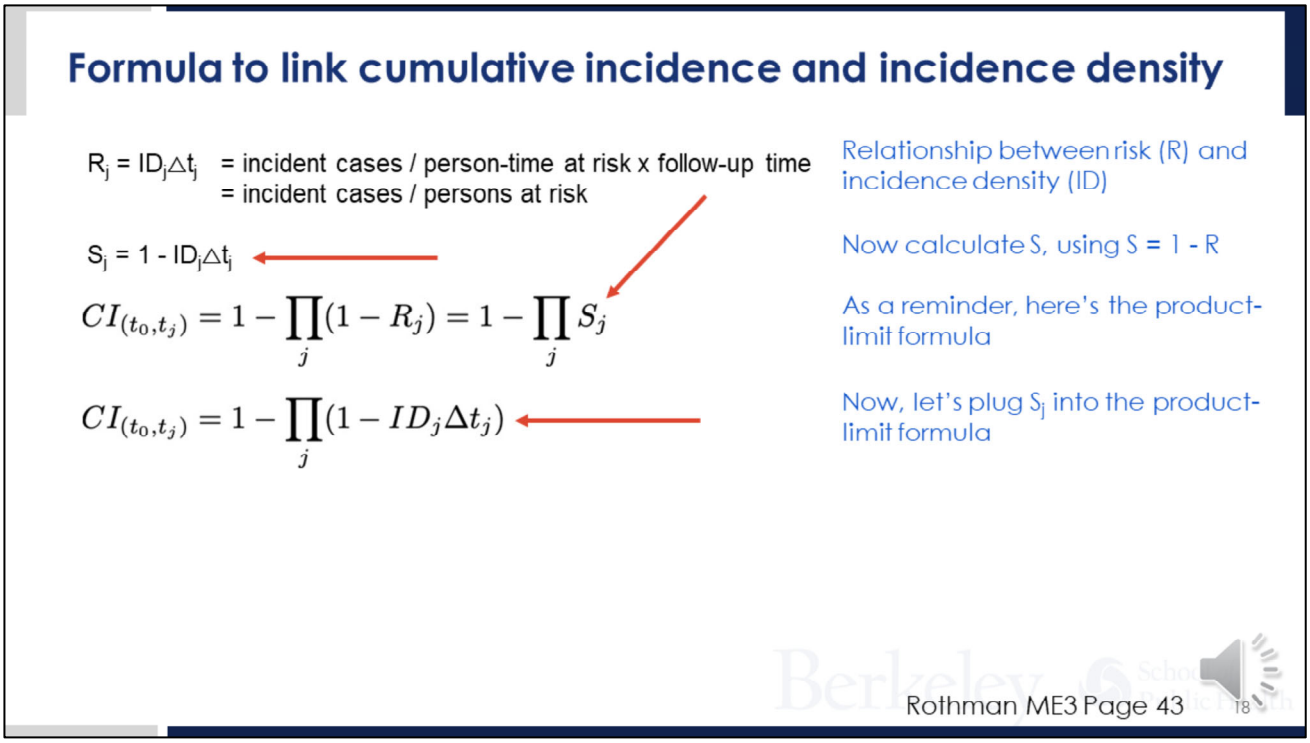
$S_j = 1 - ID_j \triangle t_j$

Now calculate S, using S = 1 - R

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

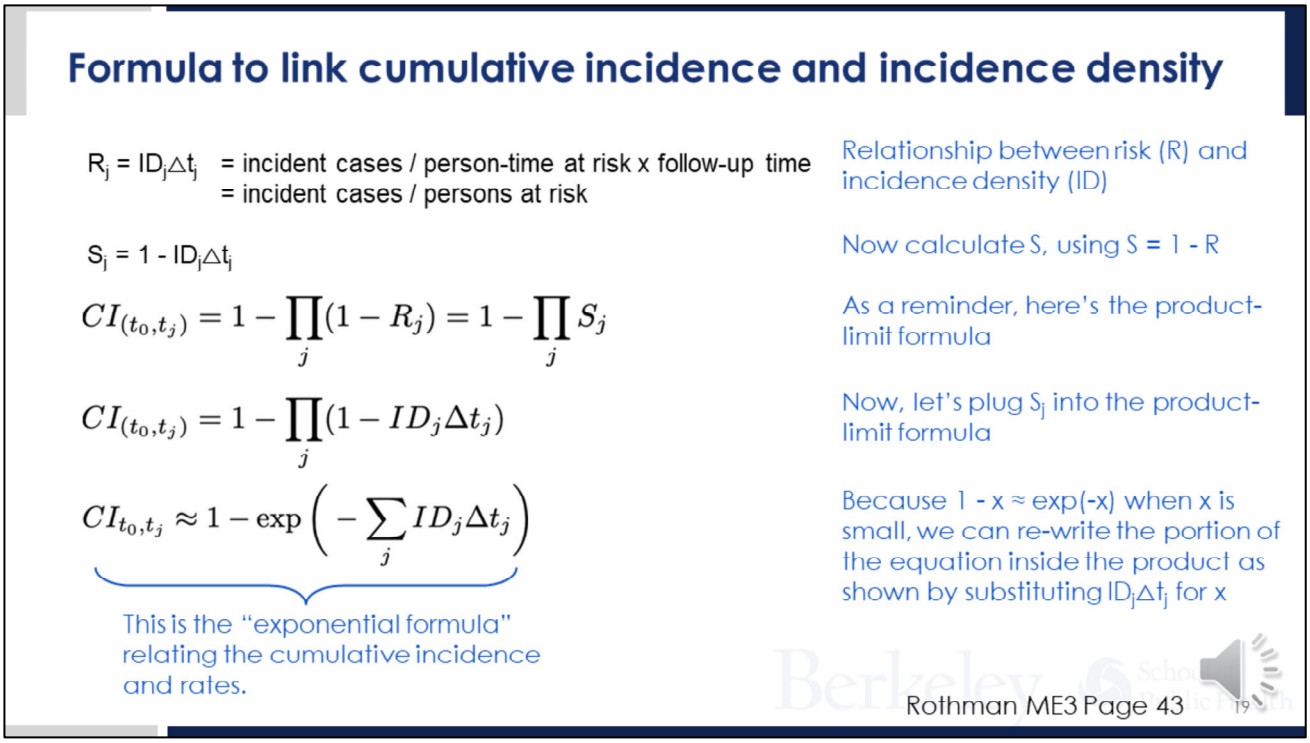
As a reminder, here's the product-limit formula

And just to remind you, here's the product-limit formula.



So in the next step, we're going to plug S sub j into the product-limit formula. So it's S sub j from step two right here is being plugged into the S sub j at the end of the product-limit formula right here.

So that gives us the cumulative incidence for the follow-up period is equal to 1 minus the grand product over all the intervals of 1 minus the incidence density in interval j times delta t.



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

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

Now, for the next step. There's a mathematical fact that 1 minus x approximate the exponent of minus x when x is small.

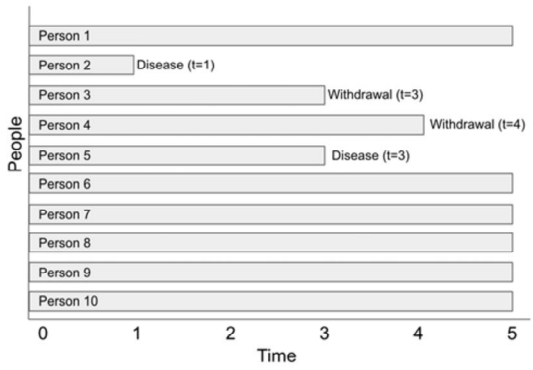
And so using this fact, we can rewrite the portion of the equation inside the product as shown to the left by substituting ID sub j delta t sub j for x in that formula. So that gives us this last formula right here. And this is referred to as the exponential formula.

And if you're really interested in this, the Kleinbaum chapter has a little bit more about why in certain circumstances under certain assumptions the rate and the risk of disease are related through an exponential formula. We're not going into all the details here. But again, that Kleinbaum chapter-- it's a great reference if you're interested in that.

So here we have it. This is our formula with our cumulative incidence on the left and our incidence density on the right. And it links these two quantities. And we're going to now use this going step by step in the density method.

Density method

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

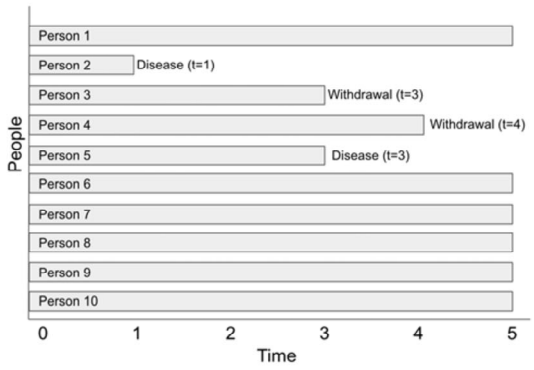


Interval j	time t_{j-1}, t_j	Pop. at risk N'_{0j}	incident disease I_j	Interval duration Δt_j	Interval incidence density ID_j	Interval risk $ID_j\Delta t_j$
1	0,1	10				
2	1,2	9				
3	2,3	9				
4	3,4	7				
5	4,5	6				

So again, here's our same example that we've used now with actuarial method and the Kaplan-Meier method and now the density method. In step one, we're going to calculate the population at risk at the beginning of each interval. So that's N prime sub 0 j. This is very similar to what we did in the actuarial method. So I'm not going to belabor this point.

Density method

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

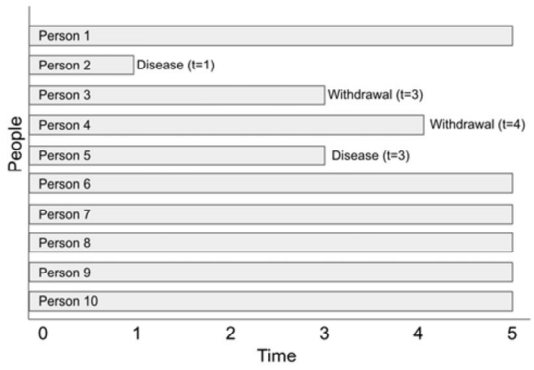


Interval j	time t_{j-1}, t_j	Pop. at risk N'_{oj}	incident disease I_j	Interval duration Δt_j	Interval incidence density ID_j	Interval risk $ID_j \Delta t_j$
1	0,1	10	1			
2	1,2	9	0			
3	2,3	9	1			
4	3,4	7	0			
5	4,5	6	0			

Similarly, we'll count the number of incident cases in each interval. Again, this is very consistent with what we did for the actuarial method.

Density method

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



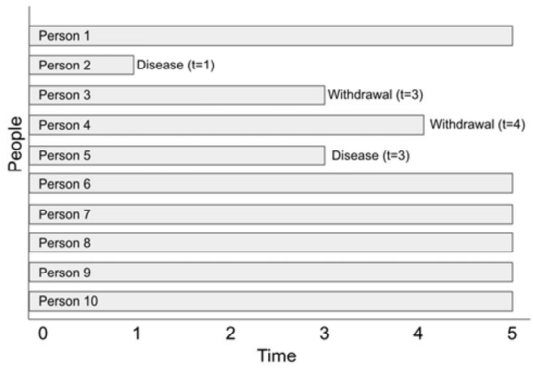
Interval j	time t_{j-1}, t_j	Pop. at risk N'_{oj}	incident disease I_j	Interval duration Δt_j	Interval incidence density ID_j	Interval risk $ID_j \Delta t_j$
1	0,1	10	1	1		
2	1,2	9	0	1		
3	2,3	9	1	1		
4	3,4	7	0	1		
5	4,5	6	0	1		

In step three, we're getting the interval duration. And we've broken up these intervals into one-time unit chunks. Like I mentioned in the previous video, we can actually use different lengths of intervals, depending on our specific research question. So this is an important column, but it actually doesn't affect anything in this particular example.

Density method

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

Keep 3 decimal points!!



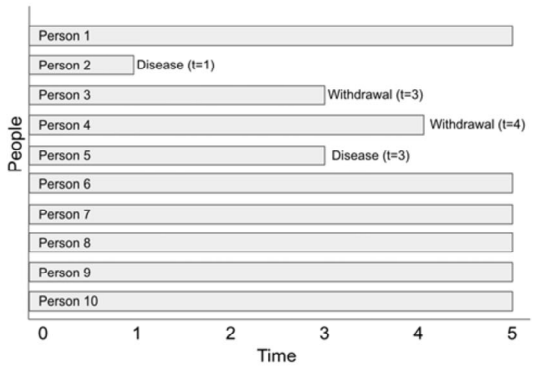
Interval j	time t_{j-1}, t_j	Pop. at risk N'_{0j}	incident disease I_j	Interval duration Δt_j	Interval incidence density ID_j	Interval risk $ID_j\Delta t_j$
1	0,1	10	1	1	1 / 10 = 0.100	
2	1,2	9	0	1	0 / 9 = 0.000	
3	2,3	9	1	1	1 / 9 = 0.111	
4	3,4	7	0	1	1 / 7 = 0.000	
5	4,5	6	0	1	0 / 6 = 0.000	

And then finally, we calculate the incidence density in each interval, which is equal to the number of incident cases in the interval divided by the population at risk times the interval duration. And so for the first interval, we have 1 over 10. And the second interval, we have 0 over 9 and so on. And because the interval duration is equal to 1, we're not really showing that here. But if there were different lengths of intervals, it'd be important to include that in the step.

Density method

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

Keep 3 decimal points!!



Interval j	time t_{j-1}, t_j	Pop. at risk N'_{oj}	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 = 0.100	0.100 * 1 = 0.100
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

Next, we're going to convert the incidence density in each interval to the risk in each interval. And so we do this by multiplying the incidence density in that interval times the duration of follow up. And again, this is all just multiplying through times 1, because the interval duration is the same in this example in each interval.

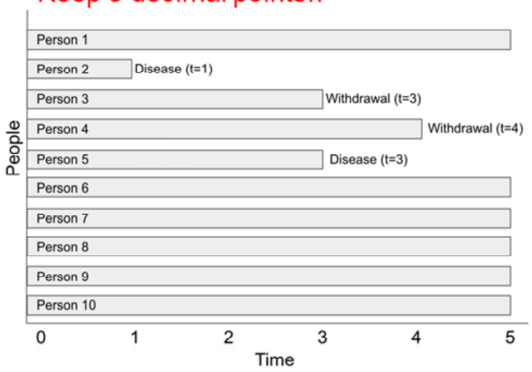
Density method

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 = 0.100	0.100 * 1 = 0.100
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(-(0.100 + 0.000 + 0.111 + 0.000 + 0.000)) = 0.190

25

And then we can use this formula, which we showed a few slides back. So this is, 1 minus the exponent of the negative sum of our interval risk, so 0.1 plus 0 plus 0.111 plus 0 plus 0. And that gives us our cumulative incidence by the density method across all the intervals. That's 0.190 when we round to three points at the very end.

Density method

- Density method estimate = 0.190
- 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.

So let's line up our different results. The density method gave us an estimate of cumulative incidence equal to 0.190. The actuarial method give us an estimate of 0.206. And the Kaplan-Meier method give us an estimate of 0.2. So the actuarial method and Kaplan-Meier methods are pretty close. And we already discussed potential reasons why they differ.

But the density method is slightly more off. And so the key thing to remember here is that the formula that we are using assumes the disease is rare. But in our example, it is not rare. Two out of five people developed the disease. And so when that's the case, that formula that approximates 1 minus x using an exponent is not going to be valid.

So when the disease is not rare, the density method using the exponential formula will not approximate the cumulative incidence and, thus, will not approximate the risk and is not appropriate to use.

Density method steps

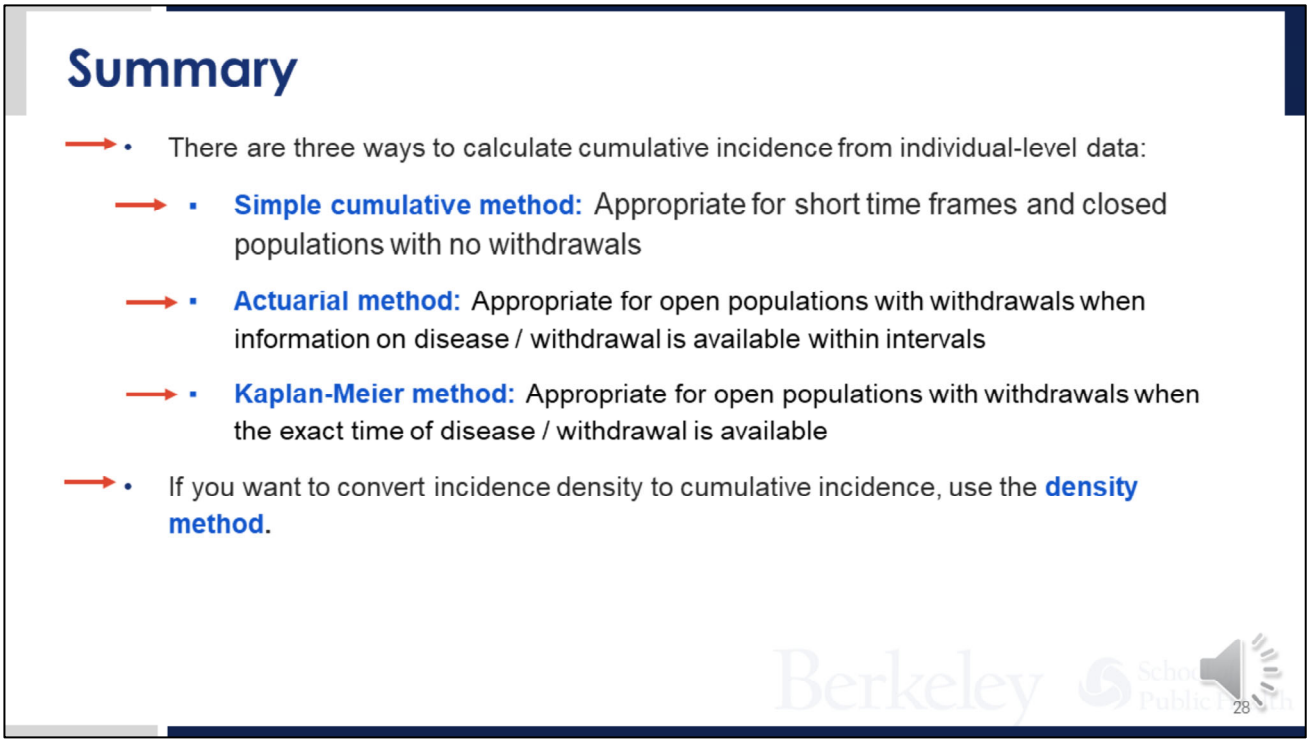
1. Step 1: Calculate the population at risk at the beginning of each interval (N'_{0j})
2. Step 2: Count the number of incident cases in each interval (I_j)
3. Step 3: Calculate the duration of follow-up in each interval (Δt_j)
4. Step 4: Calculate the incidence density in each interval: $ID_j = I_j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

27

Here's all the steps again, so you can come back to this later.



And to summarize, we've talked about three different ways to calculate cumulative incidence from individual level data. Simple cumulative method is appropriate when we have short time frames and closed populations with no withdrawals. The actuarial method is appropriate for open populations that have withdrawals when we have information about the disease occurring within intervals.

Kaplan-Meier method is also appropriate for open populations that have withdrawals when we have the exact time of the disease and withdrawal. And then if we want to convert the incidence density or rates to cumulative incidence, we can use the density method.

Let's next talk about the assumptions of cumulative incidence calculations.

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

This table shows some of the assumptions required for each of the cumulative incidence calculations that we've discussed. This can help guide which actual method to choose when making calculations in a particular problem or analysis.

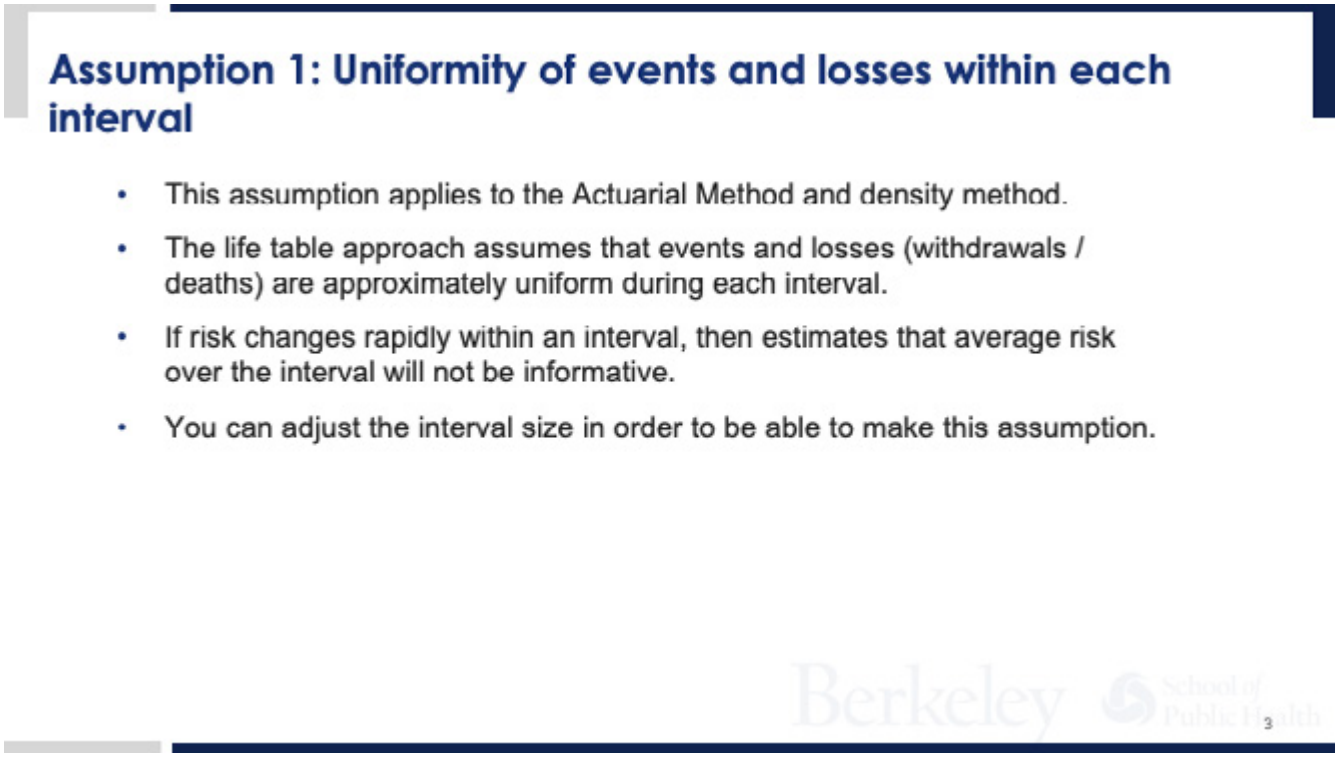
The different methods and assumptions to be aware of are first, that there is a uniformity of events and losses within each interval. So you'll recall that in the simple cumulative and actuarial methods and also in the density method, there were fixed intervals established that require this assumption.

Next, this issue of independence of censoring and survival is an assumption made across all the different methods. The following assumption is that there are no secular trends in the analysis. This applies to all four of the methods.

The next assumption is that the population is closed, and that is true for the product limit formula for the actuarial Kaplan-Meier and density method, and it's also true for the simple cumulative method. There is no account taken of competing risks. And what is meant by that is that other things are happening to the population. That's not adjusted for in any of these methods. The assumption is just made that there are no

competing risks.

And finally, that the number of events in each event time is a small proportion of the number at risk. That's not an assumption made.



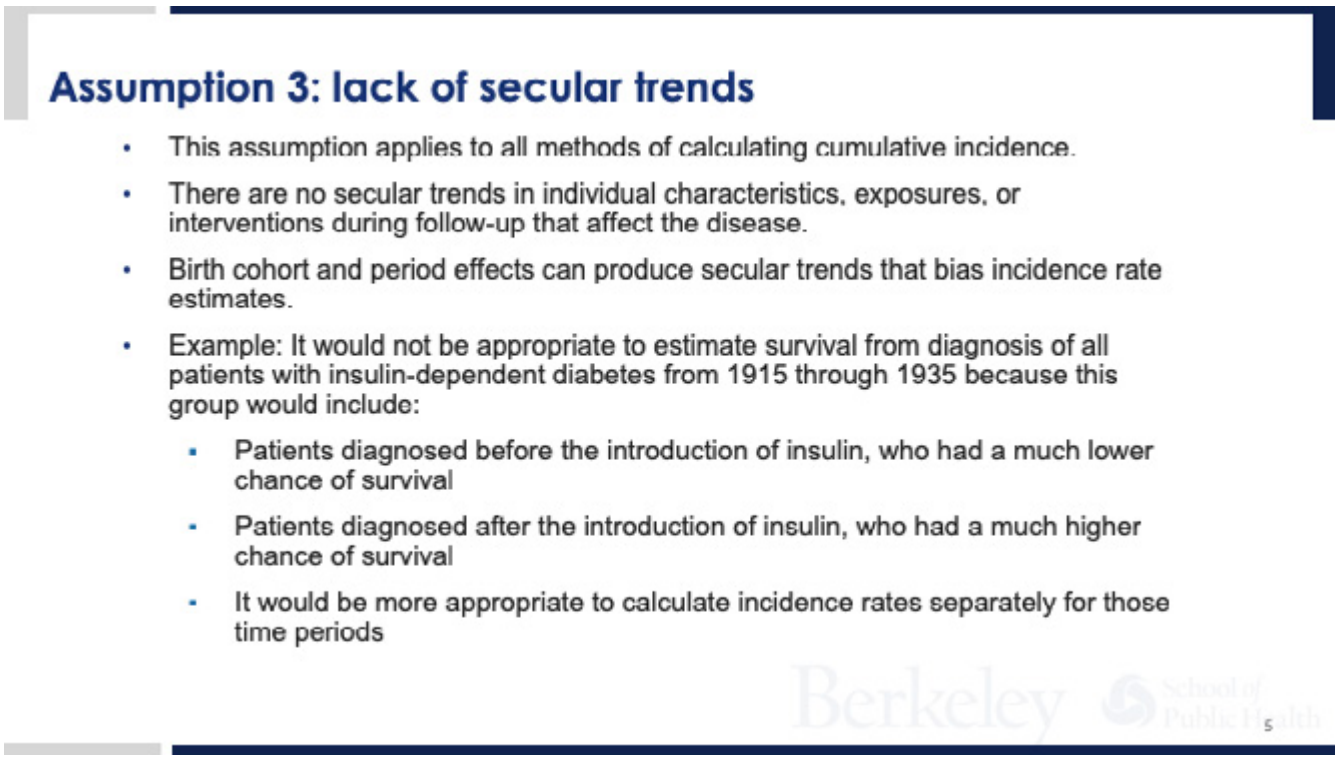
So let's talk about the first assumption. This assumption applies to the actuarial method and the density method. The life table approach assumes that events and losses-- that is, withdrawals and deaths-- are approximately uniform during each interval. If risk changes rapidly within an interval, then estimates that average risk over the interval will not be informative. You can adjust the interval size in order to make this assumption. So in other words, making the interval shorter to fit your assumption is an allowed approach to this.

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.

The second assumption is the independence of censoring and survival. This assumption applies to all methods of calculating cumulative incidence. Here, censored individuals have the same probability of the event after censoring as those remaining under observation. That is, the censoring is independent of survival.

For 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 a 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.



The third assumption is the lack of secular trends. This assumption applies to all methods of calculating cumulative incidence, and it is that 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 would bias incidence rate estimates. Here's an example.

It wouldn't be appropriate to estimate survival from the diagnosis of all patients with insulin-dependent diabetes mellitus from 1915 through 1935, because this group would include patients who were diagnosed before the introduction of insulin, who had a much lower chance of survival, as well as patients diagnosed after the introduction of insulin, who had a much higher chance of survival. Therefore, 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.

The fourth assumption is that the population is closed. This assumption applies to all methods that use the product limit formula. And this is because the incidence proportion is only defined for closed populations. In practice, though, 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.

The fifth assumption is that there are no competing risks. This assumption applies to all methods that use the product limit formula. And 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, competing risks 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 minus 1.

The population size is not constant in the interval. This assumption is made almost universally, but often it's 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)

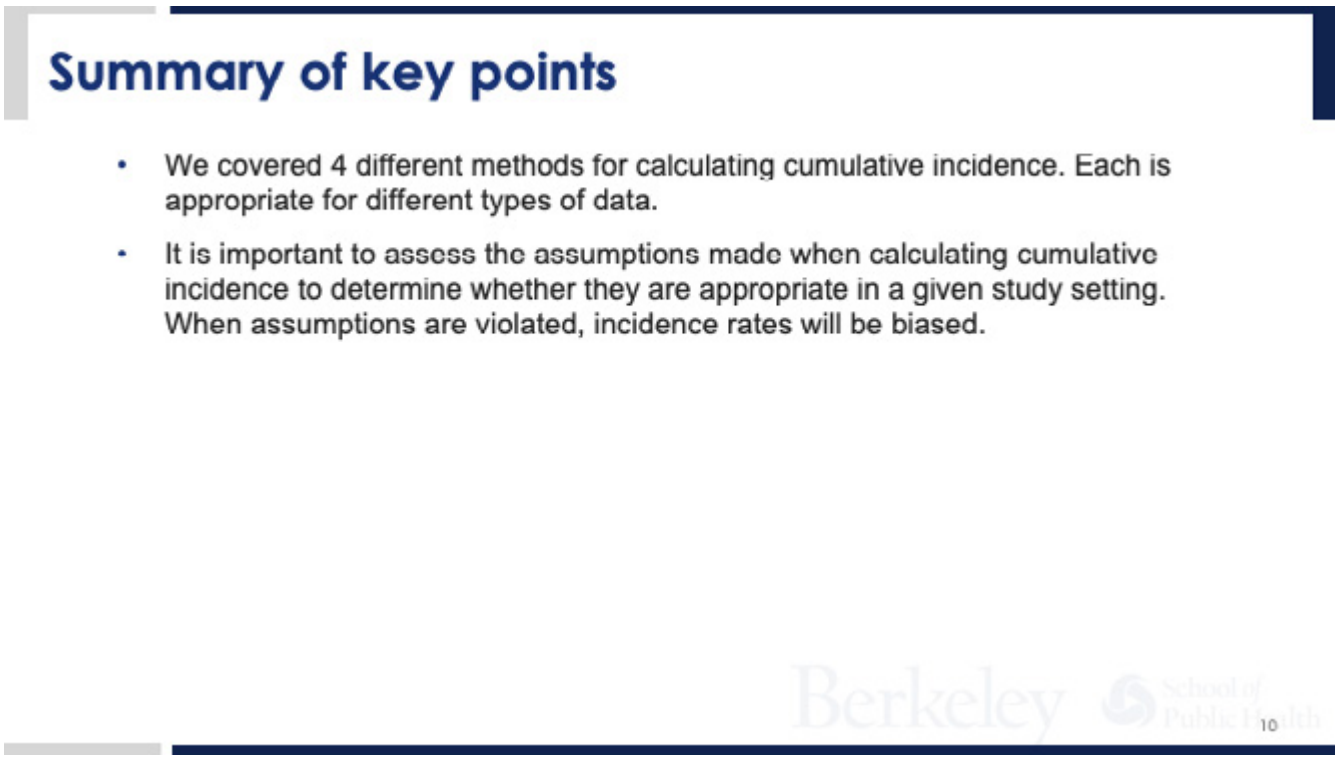
The sixth assumption is that the number of events at each event time is small in proportion to the number at risk at that time. This assumption applies to the density method, because it's required for both the product limit formula and the exponential formula that are shown below here.

Small means that the incidence rate, i times the delta t, is less than 0.1. That's a rule of thumb you can use. This assumption is required to substitute in the exponent into the product limit formula, and you can see more about this in the Rothman text.

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

So just to recap where we started, here are the assumptions required for each of the different methods we've talked about for calculating cumulative incidence.

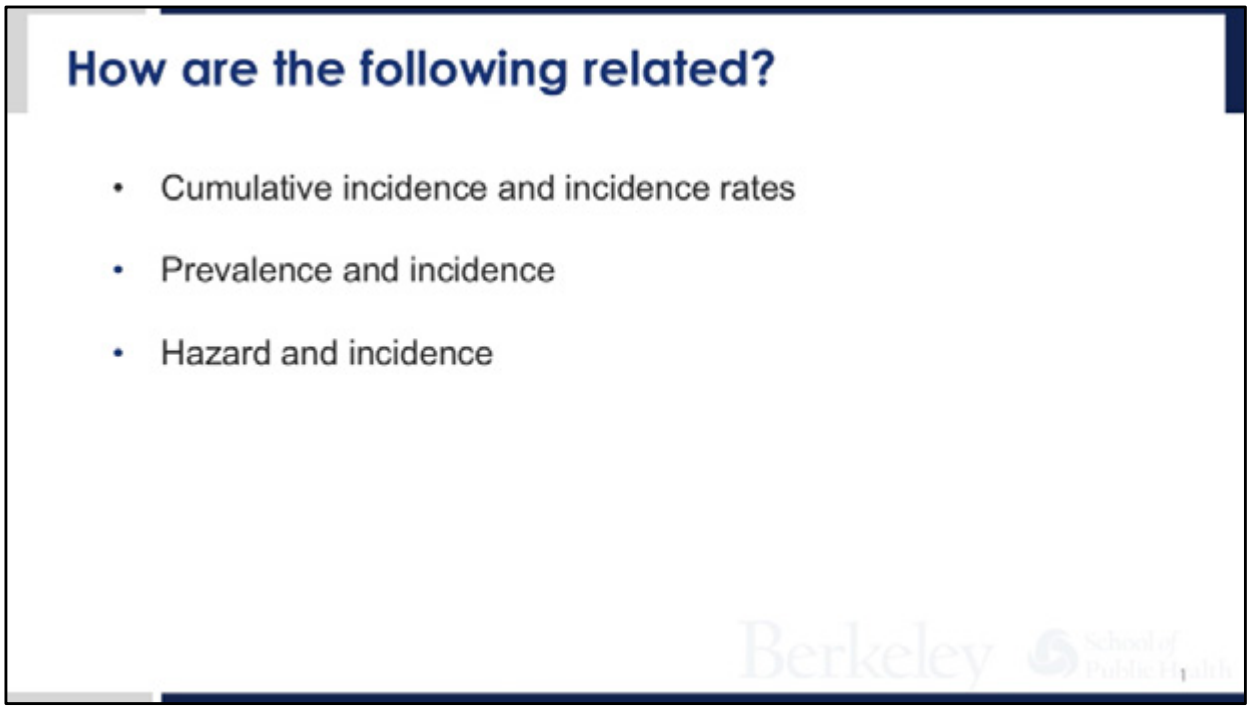


So we've covered four different methods for calculating cumulative incidence. Each is appropriate for different types of data.

It's 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 can be biased.

Lecture 3.3.1: Relationships Between Measures of Disease

In this lesson, we're going to talk about some of the interrelationships between the various measures of disease we've learned so far.



The question is, how are the following measures of disease related to each other--cumulative incidence and incidence rates, prevalence and incidence, and hazard and incidence?

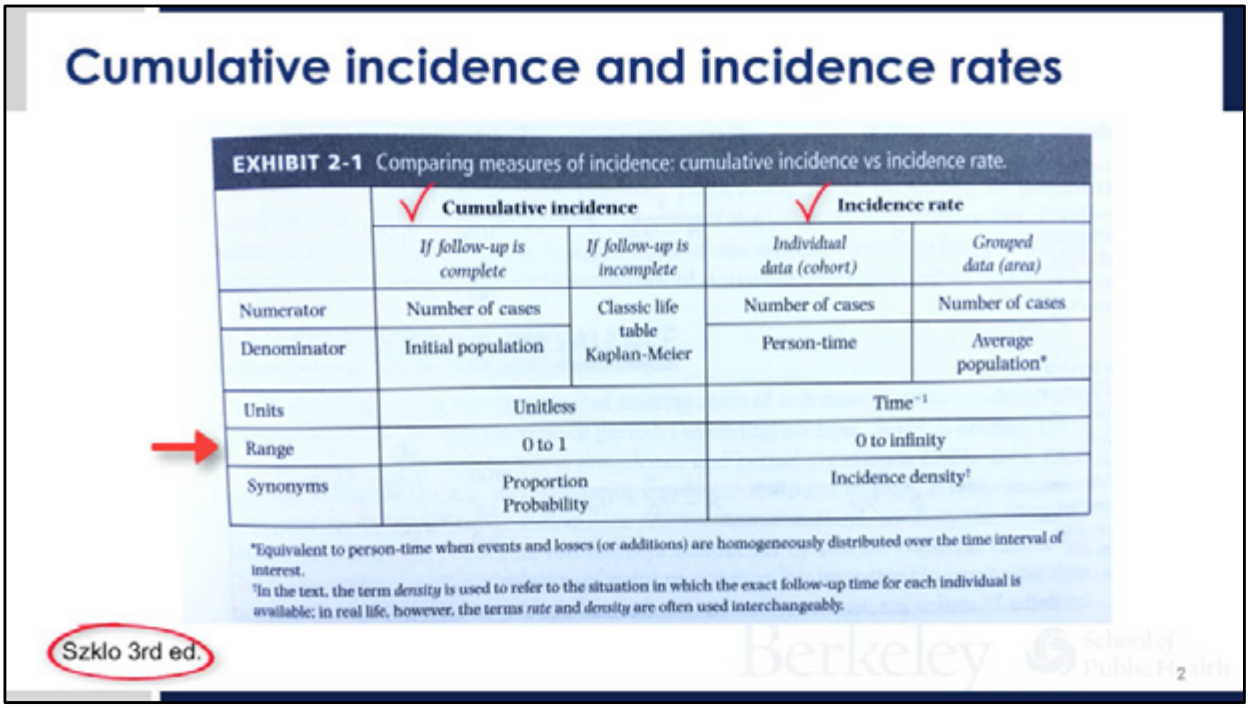


Exhibit 2.1 from the Szklo textbook, we're making some comparisons between the cumulative incidence and the incidence rate. I won't read all the particular direct comparisons to you, but what you should be able to understand is the difference between the numerator for each of these and the denominator for each of these. And then in each situation, this can differ depending on whether the follow up is complete or incomplete. Also, you should be able to understand which measures are unitless because the numerator and denominator have the same units. So the units cancel. Whereas in the incidence rate measures, time is in the denominator and time only.

They range in the case of the cumulative incidence from a low of 0 up to 1. And you should be able to understand the situation where, why can the cumulative incidence have a value of 0 and when would it have a value one? And similarly for the incidence rate, be able to understand why the incidence rate could range from a low value of 0 up to positive infinity. And finally, what are the different synonyms that are used for cumulative incidence and incidence rate.

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 = - $\Delta N / \Delta t * N'_t$
  - - ΔN 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*Δt).
  - 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

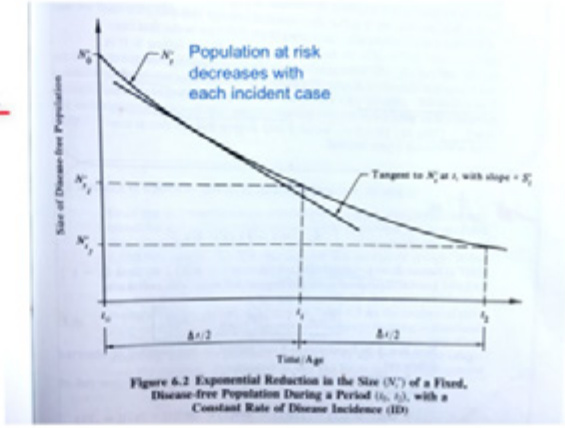


Figure 6.2 Exponential Reduction in the Size (N'_t) of a Fixed, Disease-free Population During a Period (t_0, t_2), with a Constant Rate of Disease Incidence (ID)

Kleinbaum Kupper & Morgenstern. *Epidemiologic Research*. 1982

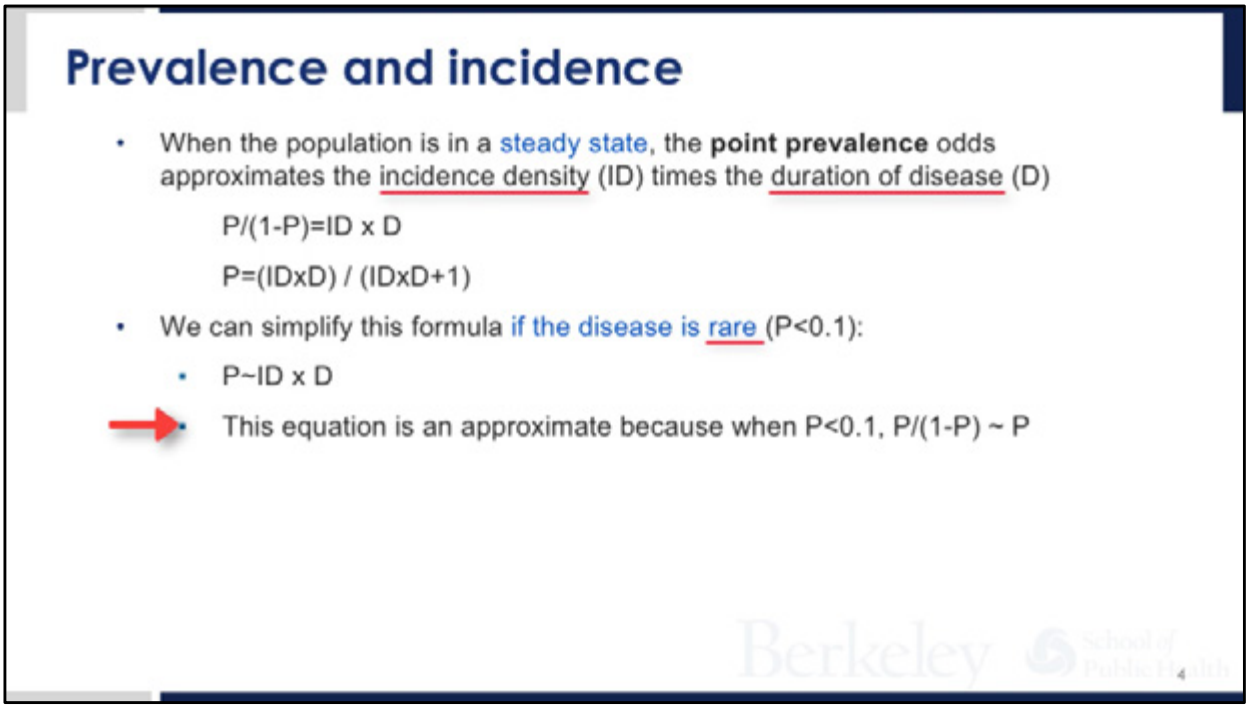
3

So the cumulative incidence and the incidence rate can be thought of sort of in graphical terms. And we can think of the incidence rate as the negative slope of the line that is touching the change in the population. So if you think of the population at risk decreasing with each incidence case on the graph, it might be plotted as a smooth curve as it decreases. And if you think back to calculus, the point at which the slope of the line intersects at that inflection point there is the tangent to the line, at some point t has a negative slope to it because the line touching the curve slopes down negatively. And that can be expressed as measured directly off the graph as the negative value of the difference between y2 and y1 divided by the difference between x2 and x1.

That's that old concept of the rise over that run, or in this case, the negative rise over the run giving us the slope of that line. And since y2 minus y1 is the change in the size of the population, that's delta n. And x2 minus x1 between two different points is the change in time. That's delta t. So the incidence rate can be expressed as the negative value for the change in n divided by the change in time, times the number of patience at the start.

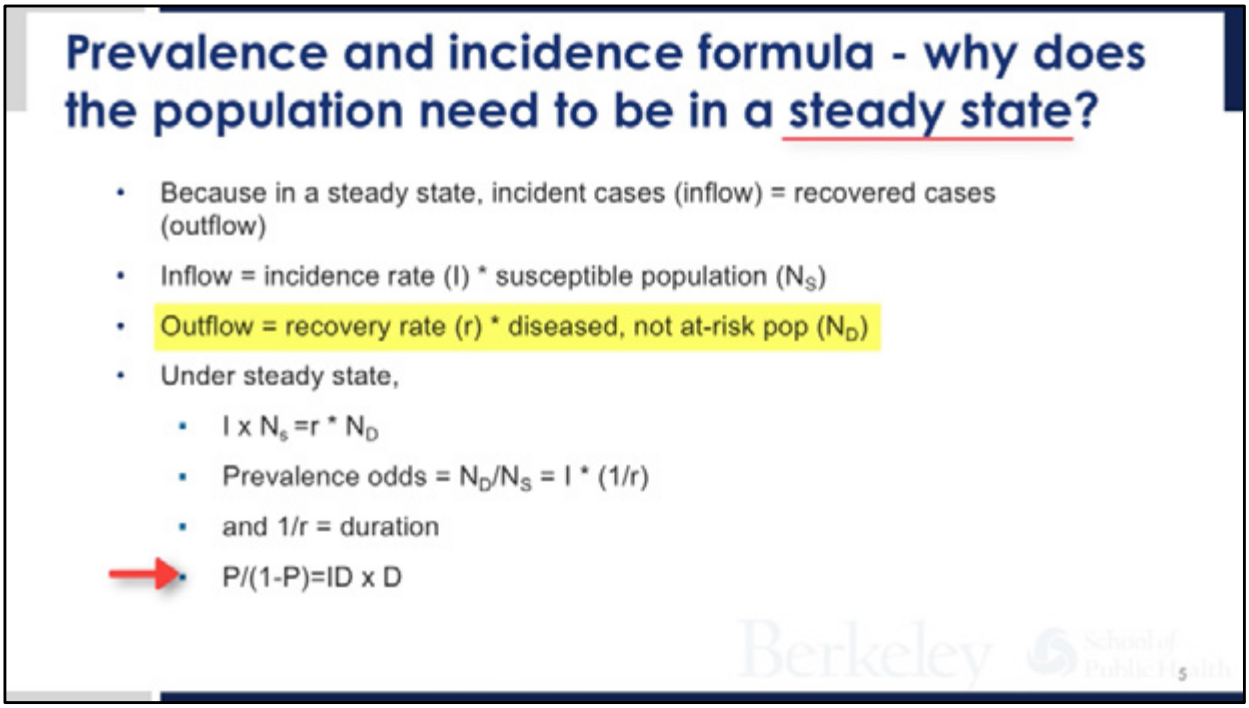
So we can think of cumulative incidence as 1 minus n prime t over n prime 0, which

can also be expressed as 1 minus this exponentiation of the negative incidence density times the delta time. We've seen that before. It's the same formula we used in the density method of calculating cumulative incidence. So the cumulative incidence is equal to 1 minus the point on a solid curve line divided by the y intercept of that line. You can take a look at the textbook for a little bit more detail on this. But this is just a kind of a graphical representation of what's happening as the population decreases in size. By that decreasing in size, I mean the population starts at a certain point, at time 0. And then as disease occurs, the disease free population smoothly reduces.



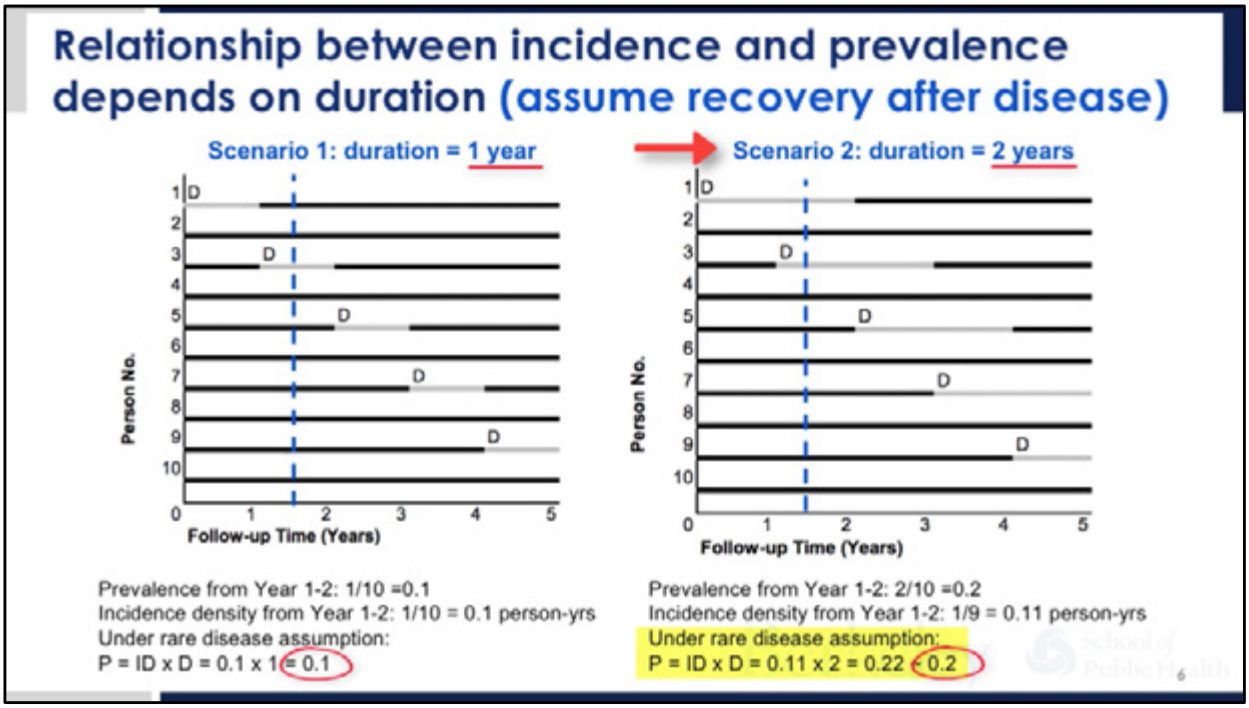
When the population is in a steady state, the point prevalence odds approximates that incidence density times the duration of diseases. This is just a formula to memorize. The probability of disease divided by the complement of the probability or 1 minus p, is equal to the incidence density times the duration of the disease. So you can see you can play with what happens as the duration, for example, of a disease gets longer.

Another way to express this is the point prevalence is the incidence density times the duration divided by the incidence density times of the duration plus 1. So if the disease is very rare, by that I mean the probability is less than 10% or less than 0.1, then the prevalence is approximately equal to the incidence density times the duration. If the duration were to become longer, the prevalence of the disease would get higher. And that simplification about the disease being rare, you can see for yourself that when p is less than 0.1, then p over 1 minus p is going to be approximately equal to p, because it will be p over 1 minus essentially 0. So p over 1 would be p.



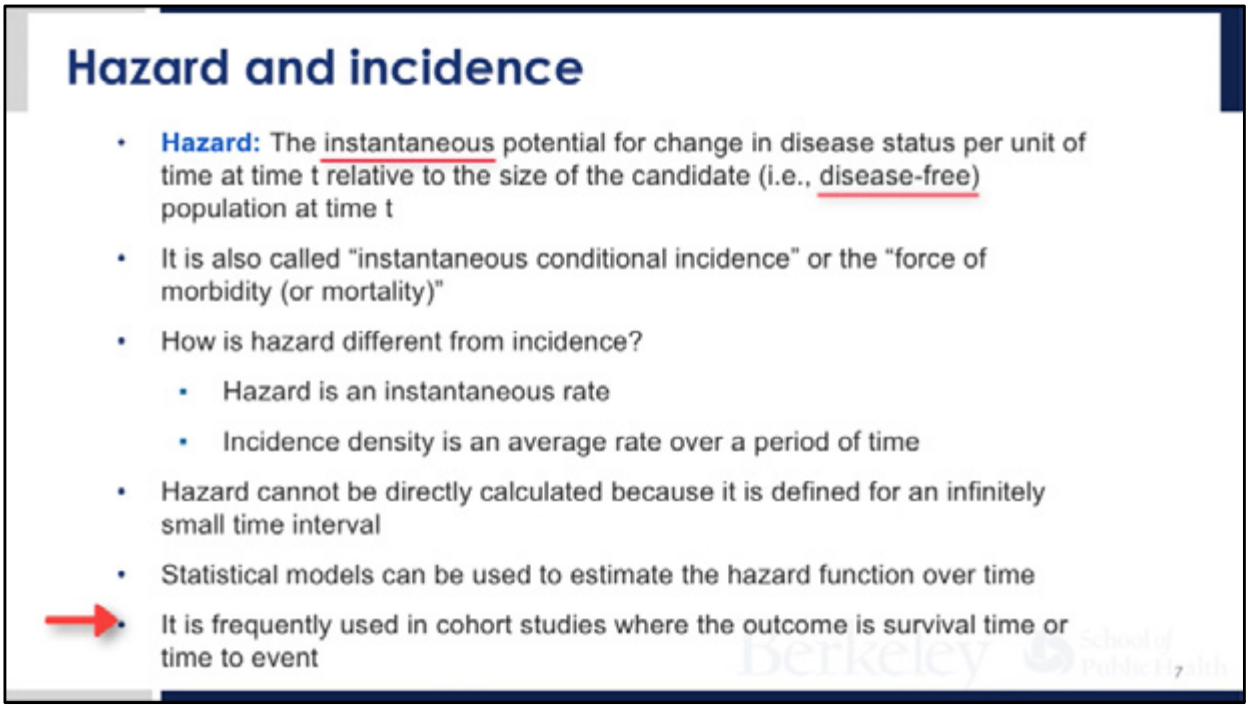
So why does the population need to be in a steady state? That's because in a steady state the incident cases-- the inflow of cases to the population equal the recovery cases or the outflow. So the incidence rate, I, is times the susceptible population N sub s is the inflow. Whereas, the outflow or the departure of patience from the population would be the recovery rate times the disease, the population not at risk.

So under steady state conditions, the incidence times the susceptible population will equal the recovery rate times the not at risk population. We can express the prevalence odds as the number with disease divided by the number susceptible. And that can be rewritten as the incidence times 1 over r. And since 1 over r equals duration, we can rewrite it is p over 1 minus p equals the incidence density times the duration, which is an algebraic re-arrangement.



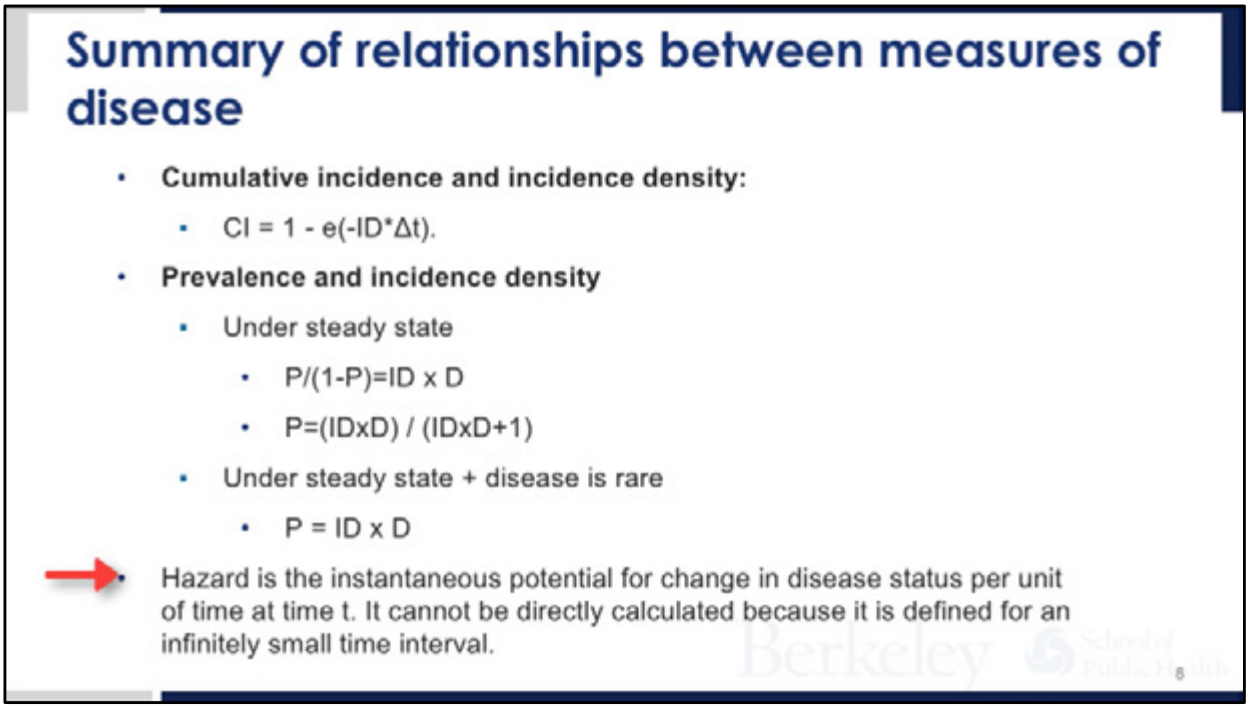
Let's look at the relationship between the incidence and prevalence and how it depends on duration. And we're assuming recovery after disease. So in scenario 1 on the left, the duration of the disease is one year. Scenario two, it'll be two years. So if we look at scenario one on the left, the prevalence from year one to two is 1 over 10 or 0.1. The incidence density from year one to two is 1 over 10 or 0.1 per person here. So under the rare disease assumption, prevalence equals the incidence density times the duration, which would be 0.1 times 1 equals 0.1.

Now let's contrast that with a situation where the duration is longer. So the point of contrast I'm trying to make here is how the duration of disease affects the prevalence. Here the prevalence from year one to two is 2 over 10 or point 0.2. The incidence density from year 1 to 2 is 1 over 9 or 0.11 per person per year. So under the rare disease assumption, the prevalence here would be the incidence density times duration, which would be 0.11 times 2, which would be 0.22 or approximately equal to 0.2.



Let's talk a little bit about a different concept, the hazard. The hazard is the instantaneous potential for change in disease status per unit of time at time t, relative to the size of the candidate population at time t. And what we mean by the candidate population, that's the disease free portion of the population. So this is also sometimes called the instantaneous conditional incidence or the force of morbidity or force of mortality, if we're talking about death.

So the question is, how is the hazard different from incidence? Well, the hazard is an instantaneous rate and the incidence density is an average rate over a period of time. So we can't calculate hazard directly because it's defined for an infinitely small time interval, and we don't have an infinitely small time interval with measurements. But statistical models can be used to estimate the hazard function over time. And the hazard is frequently used in cohort studies where the outcome is survival time or time to event.



So to summarize the relationships between the measures of disease, cumulative incidence and incidents are related by this formula where the cumulative incidence is equal to 1 minus e raised to the power negative incidence density times delta time. So think about to yourself what happens as delta time changes or what happens as negative incidence density changes to the cumulative incidence?

The prevalence and incidence density can be thought of in two situations. One, under steady state. And one under steady state when disease is rare. So under a steady state it's a little more complicated. But p over 1 minus p prevalence over 1 minus prevalence is equal to the incidence density times the density, which can be rewritten as the prevalence equals the incidence density times the density divided by the quantity and the incidence density times the density plus 1.

And under steady state where the disease is rare, the prevalence because of the simplification we discussed earlier, can equal the incidence density times the duration. So as the duration gets longer, the prevalence gets higher.

And then finally, hazard is the instantaneous potential for change in disease status per unit of time at time t. And it can't be directly calculated because it is defined for an infinitely small time interval.

AMERICAN

Journal of Epidemiology

Formerly AMERICAN JOURNAL OF HYGIENE

VOL. 102 OCTOBER, 1975 NO. 4

Reviews and Commentary

DEFINITION OF RATES: SOME REMARKS ON THEIR USE AND MISUSE

REGINA C. ELANDT-JOHNSON

In almost every scientific discipline there is a certain amount of lack of precision and ambiguity in terminology which often upsets and confuses beginners before they become accustomed to the situation and accept it (some never do).

One reason for this might be that authorities in one field borrow terminology from another field in which they are not specialists. Another reason could be a matter of semantics; a particular language may have more than one meaning for the same word; or vice versa, people use two (or more) words as synonyms though, in fact, their meanings are distinct.

The use of the word *rate* in epidemiology, demography, medicine and even in actuarial work suffers from these disadvantages; it is borrowed from physics and biochemistry and misinterpreted; it has more than one meaning in the English language; and the most common error—it is used interchangeably with the term *proportion*, because both are incorrectly assumed to be synonyms of *ratio*.

Before you proceed to the next section, please, take a piece of paper and a pencil and write down your definitions of *ratio*, *proportion* and *rate*. If you have difficulties, or are curious about what has been said about these terms in respectable books in your discipline, compare your definitions with those given by others. Finally, read the rest of this article and argue with me if you disagree with my definitions.

RATIOS, PROPORTIONS, RATES

Ratios

Ratio is, in a very broad sense, the result of dividing one quantity by another ($R = a/b$).

In sciences, however, it is mostly used in a more specific sense, that is, when the numerator and the denominator are two separate and distinct quantities; neither is included in the other. Often the quantities are measured in the same units, but this is not essential. For example,

$$\textit{Sex ratio} = \frac{\textit{(No. of males)}}{\textit{(No. of females)}}$$

$$\textit{Fetal death ratio} = \frac{\textit{(No. of fetal deaths)}}{\textit{(No. of live births)}},$$

in a given population.

An *index*, a sort of comparative summary measure of two (or more) phenomena, is often expressed as a ratio. For example,

$$\textit{Weight-height index} = kg/(cm - 100)$$

This work was supported by US National Heart and Lung Institute Contract NIH-NHLI-71-2243 from the National Institutes of Health.

Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27514.

is a ratio and it is used as a measure of obesity.

Proportions

Proportion is a type of ratio in which the numerator is included in the denominator [$p = a/(a + b)$]. For example,

[*No. of males*]/[(*No. of males*) + (*No. of females*)]

in a given community is the *proportion* of males in this community.

Epidemiologists calculate the *proportion of fetal deaths* = (No. of fetal deaths)/(No. of conceptions) and call it (incorrectly) the "fetal death rate." It is a *relative frequency* of fetal deaths among all conceptions and can be used as an estimate of the *probability* of this event.

Generally, the numerator and the denominator in $a/(a + b)$ do not need to be integers. They can be measurable quantities such as, for example, weight, length, space, volume, etc.; in such cases proportions are often also called *fractions*. For example, a mass of one part of a body can be expressed as a fraction of the total mass of the body. In random phenomena, such fractions could be used in estimating probabilities.

Concepts of instantaneous and average rates

Ratios and proportions are useful summary measures of phenomena which have occurred under certain conditions. In particular, in studies of populations, the conditions are often determined by factors such as race, sex, space, and often in a definite period of time (e.g., in a year).

On the other hand, the concept of *rate* is associated with the *rapidity of change* of phenomena such as: chemical reactions (gain or loss of mass, increase or decrease in concentration), birth, growth, death, spread of infection, etc. *per unit of time* or other variable (e.g., temperature or pressure).

Generally, a phenomenon may be described by a continuous function y of another (independent) variable x, i.e., $y = y(x)$.

Rate may be defined as a *measure of change* in one quantity (y) *per unit* of another quantity x on which y depends. Thus if $y = y(x)$ and $\Delta y = y(x + \Delta x) - y(x)$, then the *average rate of change* (i.e., the average change in y per unit of x in the interval $(x, x + \Delta x)$) is $\dfrac{\Delta y}{\Delta x}$.

Since x is usually time and y describes a continuous process in time, $\dfrac{\Delta y}{\Delta x}$ is the *average velocity* of the process. It has sign + or −, depending on whether y increases or decreases with time.

In many situations $\dfrac{\Delta y}{\Delta x}$ varies with Δx. Thus the "true" rate per unit time at the instantaneous time point x is

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \frac{dy}{dx} = \alpha(x). \qquad (1)$$

More precisely, equation 1 is the (absolute) *instantaneous rate of change in y per unit time at the time point x*. It is the *true velocity* of a process (or reaction) at time x. The "velocity curve", $\alpha(x)$, describes the pattern and direction of changes in $y(x)$.

We shall take most of our examples from life tables because they are closely related to the study of epidemiology of chronic diseases.

The basic function of the life table is the *survival function* l_x. It can be considered as a continuous monotonically decreasing function of x. The function $-\dfrac{dl_x}{dx}$, called the *curve of deaths*, describes the speed (*velocity*) of dying.

Relative rates

However, in most chemical and biologic processes not the absolute change in sub-

stance per unit of time, but the relative change per unit of time *and* per unit of substance available at this time is more informative.

For convenience, suppose that $x > 0$ is *time*, and $y(x)$ is a "mass" which is exposed to reaction (e.g., decay) at time x.

The (relative) *instantaneous rate of change per unit of mass* $y(x)$ *and per time unit at the time point* x is

$$\beta(x) = \frac{1}{y(x)} \frac{dy}{dx}. \tag{2}$$

Since this kind of rate is the most commonly used, the word "relative" is often omitted unless ambiguity might arise thereby.

In chemistry, equation 2 is often called *reaction velocity*. We notice that, in fact, equation 2 can be written as

$$\frac{d \log y(x)}{dx} = \frac{1}{y(x)} \frac{dy}{dx} = \beta(x). \tag{3}$$

If we know $\beta(x)$, we can evaluate $y(x)$. Integrating equation 3, we obtain

$$y(x) = \exp[\int_0^x \beta(u)du]. \tag{4}$$

Exponential growth model is an example. Let $P(t)$ denote the population size at time t and assume that the change in population size, $dP(t)$, is proportional to its actual size and to change in time, dt, that is

$$dP(t) = aP(t)dt,$$

where a is a growth rate. Hence

$$P(t) = P(0)e^{at} \tag{5}$$

Another example is the *force of mortality* in a life table, defined as

$$\mu_x = -\frac{1}{l_x} \frac{dl_x}{dx}. \tag{6}$$

This is the (relative instantaneous rate of change in survivorship (i.e., rate of dying) of a cohort experiencing a particular mortality pattern as described by the l_x column of a life table.

It is also possible to interpret the rate $\beta(x)$ defined in equation 2 as

$$\beta(x) = \frac{1}{y(x)} \frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\Delta y}{\int_x^{x+\Delta x} y(u)du} = \lim_{\Delta x \to 0} \frac{\Delta y}{y(x')\Delta x}, \tag{7}$$

for some x' in $(x, x + \Delta x)$.

Here the integral $\int_x^{x+\Delta x} y(u)du = y(x')\Delta x$ (shaded area in figure 1) may be interpreted as an amount of *mass-time* ("mass $\times$ time") available between x and $x+\Delta x$. We may then define a (relative) *average rate of change* of mass $y(x)$ *in* $(x, x + \Delta x)$ *per unit of time* x, *as*

$$b(x) = \frac{\Delta y}{\int_x^{x+\Delta x} y(u)du} = \frac{\Delta y}{y(x')\Delta x}, \tag{8}$$

for some x' in $(x, x + \Delta x)$.

In fact, if the mathematical form of $y(x)$ is not known, only $b(x)$ defined in equation 8 can be calculated from the data. This is why we need to use an "average rate" rather than an instantaneous one. It should be noticed that

$$p(x) = \frac{\Delta y}{y(x)} \tag{9}$$

represents that *fraction* (proportion) of the mass $y(x)$ available at time x, which has been changed (e.g., decayed) in a definite period of time x to $x + \Delta x$. This is not a rate. The rate, defined in (8), might be approximately evaluated as

$$b(x) \doteq \frac{\Delta y}{y(x) + \frac{1}{2}\Delta y}, \tag{10}$$

where the denominator represents the approximate amount of "mass-time " (Note: Δy is a negative quantity if the mass undergoes decay.)

In life tables, the *central death rate* is defined as

$$m_x = \frac{l_x - l_{x+1}}{\int_0^1 l_{x+t}dt} = \frac{d_x}{L_x}. \tag{11}$$

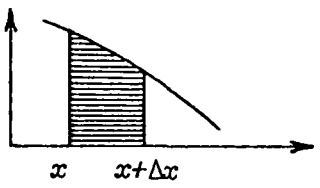


FIGURE 1.

This is a kind of "average rate." Here L_x is the total number of *person-years* lived by the population l_x in the year x to $x + 1$. We recall that in actuarial work, L_x is often approximated by $l_x - \frac{1}{2}d_x$.

The central death rate is estimated by

Age specific death rate

$$= \frac{\textit{No. of observed deaths in age } x \textit{ to } x+1}{\textit{Population at risk of this age}}.$$

Population at risk, also called in actuarial science "*central* exposed to risk," is interpreted as the equivalent number of persons, each exposed to risk of dying for a full year. In experimental populations, this number is estimated using various approximations. In calendar-year type population data, the *midyear* population size of age x last birthday is used as an approximation to the "true" size of population at risk (i.e., "central exposed to risk").

It seems that it would be useful to clarify the definition of the important life table function

$$q_x = \frac{l_x - l_{x+1}}{l_x} = \frac{d_x}{l_x}, \qquad (12)$$

which is closely related to m_x. It is sometimes called the "*mortality rate*." Clearly, it is not a rate; it is the probability (proportion) of deaths *in a year* (fixed time) but *not per year*, of those who survive to age x; the common arbitrary, though conveniently chosen time unit—one year—probably causes the confusion in using the term "rate" instead of "proportion." Also "population at risk" might be misinterpreted. The number of individuals alive at the *beginning* of the interval is sometimes called by actuaries "initial exposed to risk". Often this distinction is not observed.

Perhaps, a numerical example will make the difference between proportion and rate clearer.

Suppose that a group of 100 individuals (e.g., mice) were alive at the *beginning* of an interval of length, one year, and 40 died during this year. Then the proportion of deaths (i.e., probability of dying in this interval) is 40/100 = 0.40. Suppose that the deaths were uniformly distributed over the year. This implies that the average time of death is the midpoint of the year. Thus during the interval of one year each survivor contributed one full year to the "exposed to risk," while each individual who died contributed, on the average, only ½ of the year. The total number of "exposed to risk" is the total number of "person-years," and in our example, this is $60 \cdot 1 + 40 \cdot \frac{1}{2} = 80$. The central death rate (i.e., the average "rapidity of dying") is 40/80 = 0.50, and this is a different number from 0.40.

Incidence and prevalence

Two rather important epidemiologic terms—*incidence* and *prevalence* "rates" —may also give rise to some ambiguity. Unfortunately, there are no precise definitions of these terms.

For example, the number of *new* cases of a disease which occurs per year in a given community is the *absolute incidence rate;* it does not refer to the population size (Note: the number of deaths per year might be called the "absolute death incidence rate," but it is not customary to use this expression.)

To calculate relative rates, however, we must evaluate the "person-periods" (often person-years) or obtain an estimate of them. If the period is a year, the size of the *midyear* population is used as an estimate of person-years exposure and the relative *incidence rate* is the number of new cases per person per year (or per 1,000 persons per year).

If, however, we calculate the ratio of the number of new cases of a disease to the number of individuals free of the disease at the beginning of the time interval, then this will be a proportion, not a rate. For epidemics, when the development of the disease is very rapid, these two quantities may differ considerably. The proportion can still be called "incidence" if one wishes (perhaps "probability of incidence" would be better), but not a rate.

In contrast, *prevalence*, which is the ratio of the number of cases at a given time to the size of the population at that time is clearly always a proportion. The term prevalence is quite legitimate, but "prevalence rate" is an impossible concept.

Conclusion

In summary, we may say that in sciences, ratios, proportions and rates are precisely defined and cannot be used as synonyms. Ratios are used as indices, proportions are relative frequencies or fractions and often estimate (or are) probabilities of certain events, while rates describe the velocity and direction (patterns) of changes in dynamic processes.

After writing this article, I gave some thought to the question: what *proportion* of research workers in disciplines mentioned at the beginning of this paper will be "converted" and use these definitions, and what might be the "*rate of conversion*". I guess, both will probably be rather low since training and long-established habits may have stronger effect than the "catalytic" power of mathematical arguments.

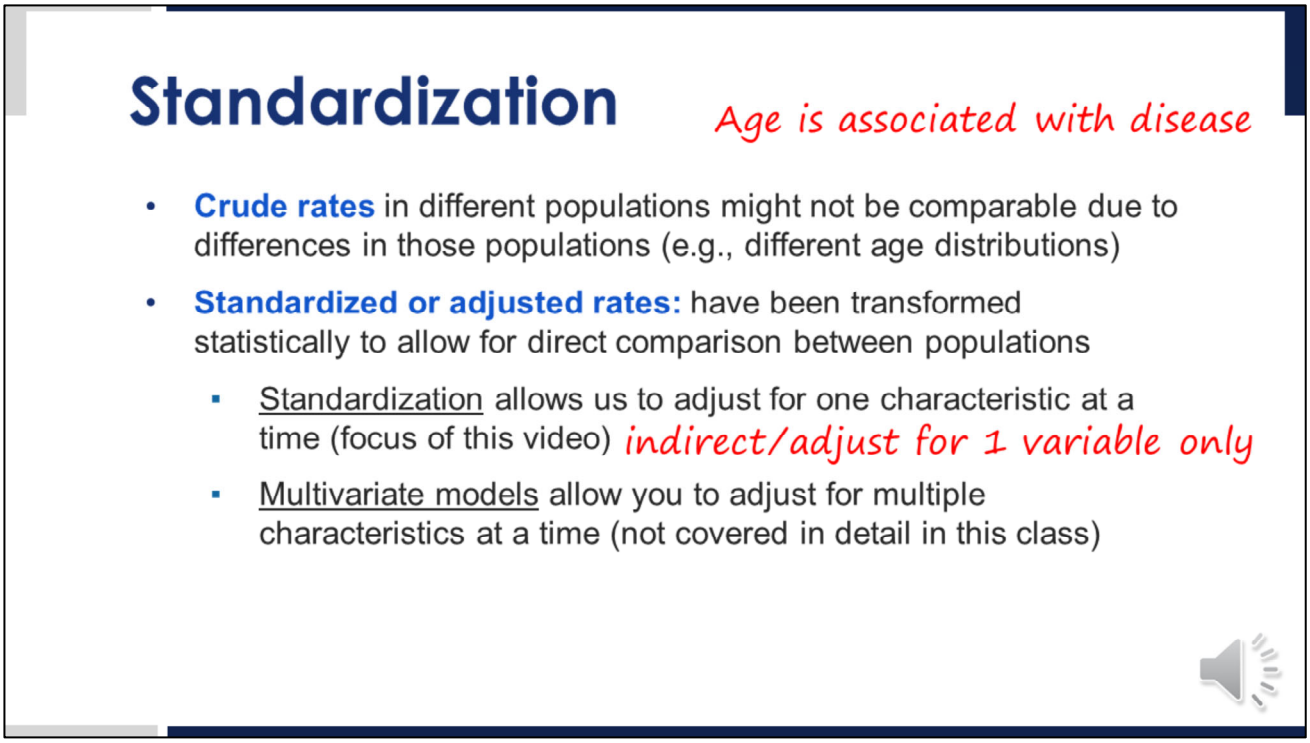
But *at any rate* (*!*), I have tried my best to clarify the distinction between ratios, proportions and rates.

Editor's note.—We suspect that most epidemiologists occasionally or regularly misuse one or another of the terms so carefully defined by Dr. Elandt-Johnson. We share her pessimism regarding the likelihood of changing this situation.

We formerly took care to refer only to prevalence *ratios* (never rates—well, hardly ever), but even this expression she indicates is incorrect, the proper term being just prevalence or perhaps prevalence proportions. Alas, so general is the term prevalence rate that many who are aware of this distinction nevertheless use it. Thus, Sir Austin Bradford Hill in the seventh edition of his widely-used and excellent textbook, *Principles of Medical Statistics*, does so.

Lecture 3.4.1: Indirect Standardization

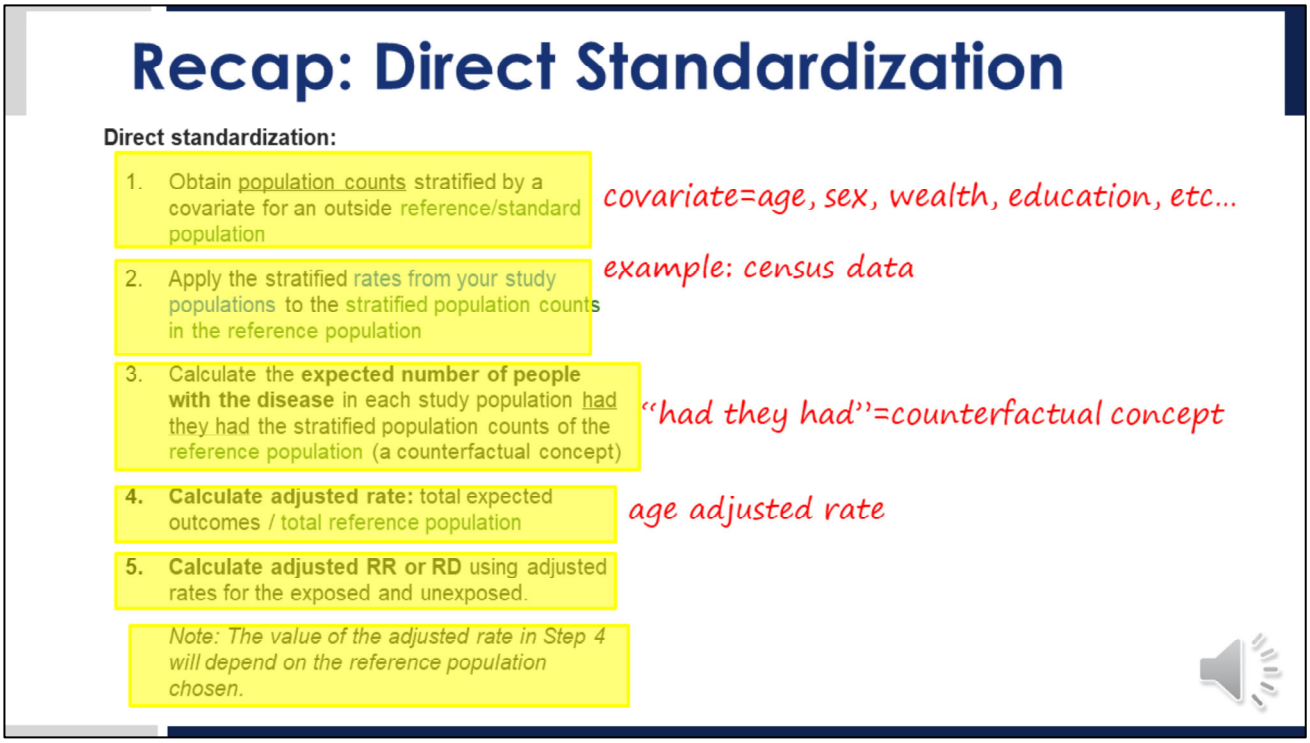
In this video, we'll learn about indirect standardization.



First let's recap standardization in general. We typically start an analysis with crude rates from different populations. And if we want to make comparisons of these rates between populations, we have to be really careful about the distribution of different covariate or confounders or other variables such as age between these populations.

For example, if the age of two populations is very different, we might not be able to make direct comparisons of crude rates between these populations, *because age is typically associated with disease. This is true not just for age, but really any variable associated with our outcome or disease of interest. One way to be able to make comparisons between populations is to calculate standardized or adjusted rates. So these are rates that have essentially just been transformed statistically to allow us to make these direct comparisons.

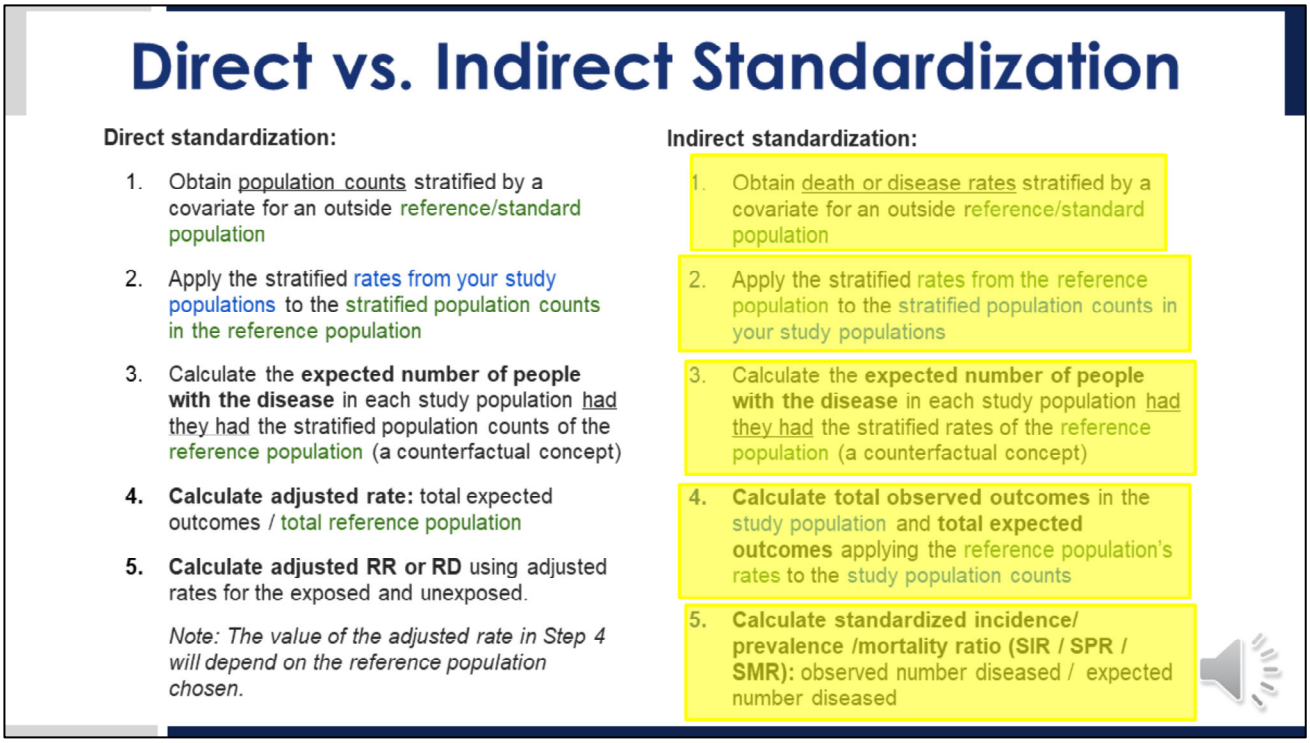
And this video is going to focus on standardization and in particular *indirect standardization. Hopefully you've already become familiar with direct standardization. I just want to briefly note that standardization *allows us to adjust for one characteristic or one covariate or confounder at a time. We can only standardize for age or only for sex, etc. If we're concerned about multiple variables being different between our populations of interest, we need to use multivariate statistical models. And that's not something that we cover in detail in this class. But if you're interested in that, take an advanced statistics class.



Let's recap the steps involved in direct standardization. *We start by obtaining population counts stratified by a covariate of interest for an outside reference or standard population. So what do we mean by covariate? *Well, we mean a variable such as age or potentially sex or wealth or education that's strongly associated with our disease of interest.

*We then apply the stratified rates from our study population to the stratified population counts in the reference population. So a common example of this would be to take *census data with age stratified population counts and use that as our reference. We then calculate the expected number of people with the disease in each study population had they had the stratified population counts of the reference population. So the had they had part of that sentence signals to us that this is a counterfactual concept. We don't know what the expected number of people with the disease would have ultimately been if our study population had the same population counts as the reference. And so that's why this is a counterfactual concept.

We then calculate an adjusted rate, which is the total expected diseases or death cases divided by the total reference population. For example, if we used age as our covariate, we would call it our age-adjusted rate. And then if we want to, we can calculate an adjusted measure of association such as relative risk or risk difference, comparing our adjusted rates for the exposed and the unexposed. And it's important to just briefly note that the values for adjusted rates we calculate in step four are going to depend on the reference population that we choose. We have to be careful about this when we make interpretations of our results.



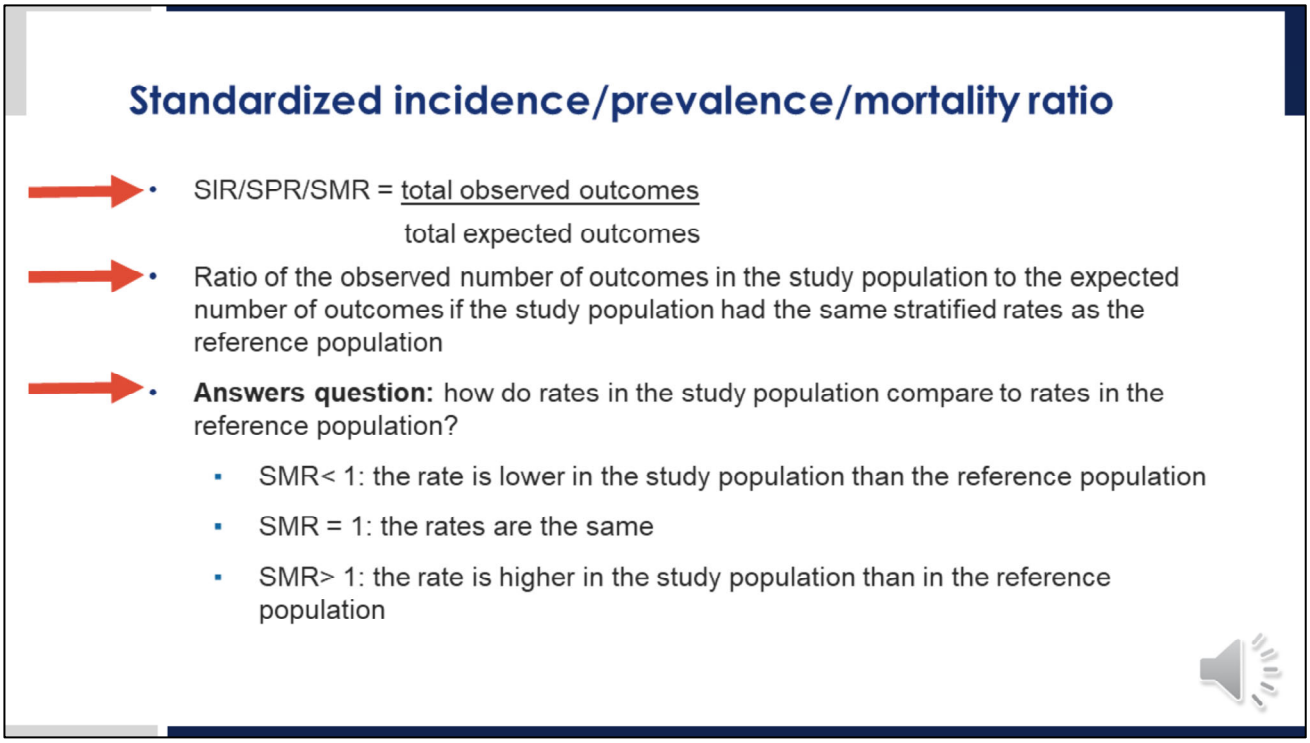
Now let's contrast that with the steps involved in indirect standardization. *In indirect standardization, we start with our death or disease rates stratified by a covariate like age from an outside reference or standard population. Direct standardization is taking population counts from the reference. Indirect standardization is taking rates from the reference.

*We then apply the stratified rates from the reference population to stratified population counts from our study. So just looking here at number two in each of these lists, we see that we're using the same information in the calculations, just from different sources in direct and indirect standardization.

*We then calculate the expected number of people with the disease or the expected number who passed away in each study population had they had the stratified rates of the reference population. Again, that is a counterfactual concept.

*And then in step four, we calculate the total observed outcomes in the study population and the total expected outcomes from the reference population when we apply the reference population rates to the study population counts.

*And then we calculate something called a standardized mortality ratio. It can also be called a standardized incidence ratio or standardized prevalence ratio. But the most common term we'll see is SMR, standardized mortality ratio.
Notice that these two processes end with different measures. And we're going to come back to why that is. But I just want to flag that for now.



What is this standardized mortality ratio? *It's the total observed outcomes divided by the total expected outcomes. *And this is essentially the ratio of observed number of outcomes in the study population to the expected number if the study population had the same stratified rates as the reference population. This measure, the SMR,* is essentially answering the question: “How do rates in the study population compare to rates in the reference population? “This is different from how we usually think as epidemiologists.

We often think of how does the rate in the exposed group compare to the rate in the unexposed group? And this is not that. This is making a comparison between the study population and the reference.

If our SMR is less than 1, it means the rate is lower in the study population than in the reference. If it's equal to 1, the rates are the same in the two populations. And if it's greater than 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

Let's go over a numerical example. Here we have two study groups, A and B. And we have N, which is the population counts.

We have deaths. And we have age stratified rates in our population. And our two age groups of interest are under 40 years and over 40 years. Our standard population rates are 12% for those who are under 40 and 50% for those who are over 40. So that's step one, obtain rates from the standard population.

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 two is to apply the standard population rates to the study populations. To do that, we're going to take our *12% for under 40 and multiply that times 100, *which is the population count in the under 40 group in study group A. *We'll do the same for the second age group in study group A.

So we'll take 50% from the standard population and multiply that times 500, which is the N for over 40 in study group A. *And then we do something analogous for study group B. Here we have 12% times the N for under 40 in study group B 500. Then we do the same for over 40.

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 three is to calculate the expected number of deaths in each study group. So that just means we need to multiply through in these cells here. So 100 times 12% equals 12. 500 times 50% is 250.

We take the sum of those two things. And that gives us 262, the total number of deaths in study group A. And we do the same thing in study group B, which gives us 110.

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

Step four is to calculate the standardized mortality ratio. We take the total observed number of deaths here, which was 110. And we divide that by the expected deaths, which is 262 for study group A. And that gives us an SMR of 0.42. And then we do the same thing for study group B. 70 total deaths that were observed in study group B divided by the total expected deaths gives us 0.64.

How do we interpret these? Well, for study group A, we can say that the mortality rate in study group A is lower than in the standard population. And for group B, the mortality rate in study group B is lower than in the standard population as well.

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.

Now in the previous slide, you may have wanted to say that study group A had a lower mortality rate than study group B because the SMR was lower, but we need to be really careful about making comparisons of SMRs. When we make a comparison using more than one study population, we're essentially using different standards. And this is because in indirect standardization, the true standard is actually our own study population. It's those population counts stratified by our variable of interest. As a result, it's usually *not appropriate to compare SMRs, SPRs, SIRs between different study groups.

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	

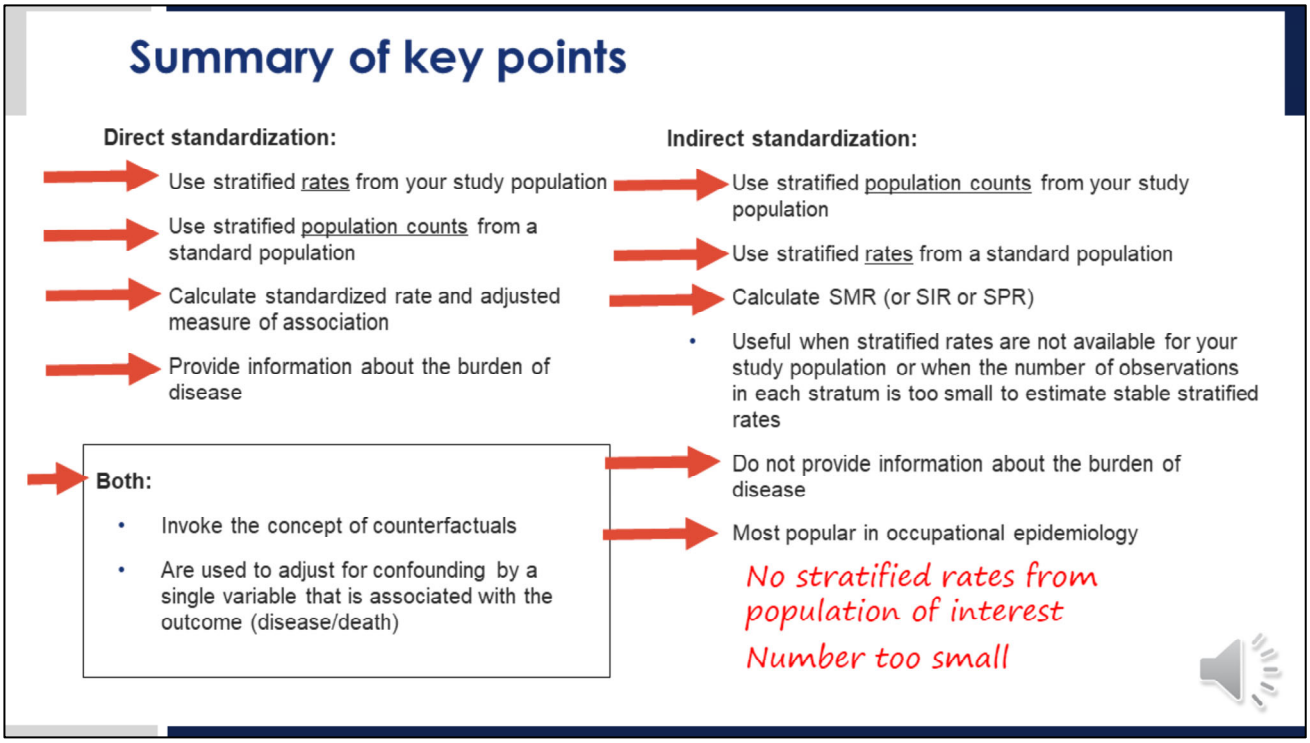
Interpret on its own!

- In this example, two study groups have identical age-specific rates, but their age distributions are different. (opposite) / different standards
- 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

So why exactly can we not we compare SMRs for group A and B? Well, if we look at the age distribution in group A and B, we can see that *it's completely different. In fact, it's *opposite. In group A, *we have far more people over 40. And in group B, *we have far more people under 40.

Because these two age stratified populations are so different between group A and B, we're *essentially using completely different standards when we calculate our SMRs for each study group. We don't want to compare 0.42 to 0.64. *Instead, we just want to interpret 0.42 on its own and then separately interpret 0.64 on its own.



To summarize what we've learned here, direct standardization uses stratified rates from your study population, whereas indirect uses stratified population counts from your study population. Direct standardization uses stratified population counts from a standard population while indirect standardization uses stratified rates from a standard population. And these two methods have different measures that they're leading to. So for direct standardization, we're ultimately going to calculate a standardized rate and an adjusted measure of association, whereas for indirect standardization our goal is to calculate an SMR, standardized mortality ratio or a standardized incidence ratio or standardized prevalence ratio.

Direct standardization provides us with information about the burden of disease. Unfortunately, indirect standardization does not, because its key comparison is between the study population and the reference population of interest. And the disease rates are coming from the reference population instead of from the study population.

So then why would we use indirect standardization? *Well, it turns out that it's quite popular in occupational epidemiology because it's *often not possible in the types of studies that are done in that subfield to get stratified rates from the study population of interest, or *sometimes the number of observations in each stratum are just too small to do a good job estimating stratified rates. And so when that's the case, it's better to use indirect standardization.

Most of the time, we prefer to use direct standardization if we're able to. And then both types of standardization invoke the concept of counterfactuals and are a method that are used to adjust for confounding by a single variable. They cannot be used to adjust for multiple variables at a time.

American Journal of Epidemiology

Vol. 186, No. 7
DOI: 10.1093/aje/kwx019
Advance Access publication:
May 11, 2017

Original Contribution

Acute Illness Among Surfers After Exposure to Seawater in Dry- and Wet-Weather Conditions

Benjamin F. Arnold*, Kenneth C. Schiff, Ayse Ercumen, Jade Benjamin-Chung, Joshua A. Steele, John F. Griffith, Steven J. Steinberg, Paul Smith, Charles D. McGee, Richard Wilson, Chad Nelsen, Stephen B. Weisberg, and John M. Colford, Jr.

* Correspondence to Dr. Benjamin F. Arnold, Division of Epidemiology, School of Public Health, University of California, Berkeley, 101 Haviland Hall, MC #7358, Berkeley, CA 94720-7358 (e-mail: benarnold@berkeley.edu).

Initially submitted September 8, 2016; accepted for publication January 23, 2017.

Rainstorms increase levels of fecal indicator bacteria in urban coastal waters, but it is unknown whether exposure to seawater after rainstorms increases rates of acute illness. Our objective was to provide the first estimates of rates of acute illness after seawater exposure during both dry- and wet-weather periods and to determine the relationship between levels of indicator bacteria and illness among surfers, a population with a high potential for exposure after rain. We enrolled 654 surfers in San Diego, California, and followed them longitudinally during the 2013–2014 and 2014–2015 winters (33,377 days of observation, 10,081 surf sessions). We measured daily surf activities and illness symptoms (gastrointestinal illness, sinus infections, ear infections, infected wounds). Compared with no exposure, exposure to seawater during dry weather increased incidence rates of all outcomes (e.g., for earache or infection, adjusted incidence rate ratio (IRR) = 1.86, 95% confidence interval (CI): 1.27, 2.71; for infected wounds, IRR = 3.04, 95% CI: 1.54, 5.98); exposure during wet weather further increased rates (e.g., for earache or infection, IRR = 3.28, 95% CI: 1.95, 5.51; for infected wounds, IRR = 4.96, 95% CI: 2.18, 11.29). Fecal indicator bacteria measured in seawater (*Enterococcus* species, fecal coliforms, total coliforms) were strongly associated with incident illness only during wet weather. Urban coastal seawater exposure increases the incidence rates of many acute illnesses among surfers, with higher incidence rates after rainstorms.

diarrhea; *Enterococcus*; rain; seawater; waterborne diseases; wound infection

Abbreviations: CI, confidence interval; IRR, incidence rate ratio.

Freshwater runoff after rainstorms increases levels of fecal indicator bacteria measured in seawater (1), but little is known about whether persons who participate in ocean recreation have a higher risk of acute illness after rainstorms. Absent epidemiologic studies to inform beach management guidelines after rainstorms, California beach managers post advisories at beaches that discourage contact with seawater for 72 hours after rainfall—a practice that is based on fecal indicator bacteria profiles in storm water outflows, which typically decline to prerainstorm levels within 3–5 days (2, 3).

In prospective cohorts in California, investigators have found increased incidence of gastrointestinal illness and other acute symptoms (e.g., eye and ear infections) associated with seawater exposure during dry summer months (4–8). In the same studies, researchers found that levels of fecal indicator bacteria in seawater were positively associated with incident gastrointestinal illness if there was a well-defined source of human fecal contamination impacting the seawater (4–8). Individual cases of acute infections and deaths associated with waterborne pathogens have been reported among surfers in southern California who surfed during or after rainstorms (9), and 2 cross-sectional studies of surfers found that seawater exposure after heavy rainfall increased reported illness (10, 11). To our knowledge, there have been no prospective studies to determine whether rainstorms increase illness among persons who participate in ocean recreation and no studies that have evaluated whether levels of fecal indicator bacteria are associated with incident illness during wet weather periods.

We conducted a longitudinal cohort study among surfers in San Diego, California. We focused on surfers because they are a well-defined population that regularly enters the ocean year-round, even during and immediately after rainstorms, given that surfing conditions often improve during storms (12). Our objectives were to determine whether exposure to seawater increased rates of incident illness among surfers compared with periods when they did not surf in order to determine whether exposure during or immediately after rainstorms increased rates more than did exposure during dry weather. We also sought to evaluate the relationship between levels of fecal indicator bacteria in seawater and incident illness rates during dry and wet weather.

METHODS

Setting

Southern California has one of the most urbanized coastlines in the world, and it receives nearly all of its annual rainfall during the winter months (November–April). San Diego County beaches have some of the best water quality in California based on levels of fecal indicator bacteria, but water quality deteriorates after rainstorms (13). The most heavily used beaches in the region are affected by urban runoff after storms, and local beach managers post advisories that discourage water contact within 72 hours of rainfall. In the present study, we focused enrollment and conducted extensive water quality measurement at 2 monitored beaches within San Diego city limits—Ocean Beach and Tourmaline Surfing Park. Both monitored beaches have storm-impacted drainage, attract surfers year-round, and have water quality levels similar to those of other beaches in the county (13). Ocean Beach is adjacent to the San Diego river, which drains a 1,088-km^2 varied land-use watershed with many flow-control structures; Tourmaline Surfing Park is adjacent to Tourmaline Creek and a storm drain, which together drain an urban, largely impervious, 6-km^2 watershed (Figure 1). The study's technical report includes additional details (14).

Study design and enrollment

We conducted a longitudinal cohort study of surfers recruited in San Diego over 2 winters, with enrollment and follow-up periods chosen to capture most rainfall events in the region. During the first winter (open enrollment from January 14, 2014, to March 18, 2014; end of follow-up on June 4, 2014), we enrolled surfers through in-person interviews at the 2 monitored beaches and through targeted online advertising on Surfline.com, a popular website on which surf conditions are reported. We enrolled participants at monitored beaches and online to assess whether individuals enrolled through these 2 modes were similar in their exposures and other characteristics. Participants enrolled on the beach were very similar to those enrolled online (Table 1), so we exclusively enrolled participants through the study's website during the second winter (open enrollment from December 1, 2014,

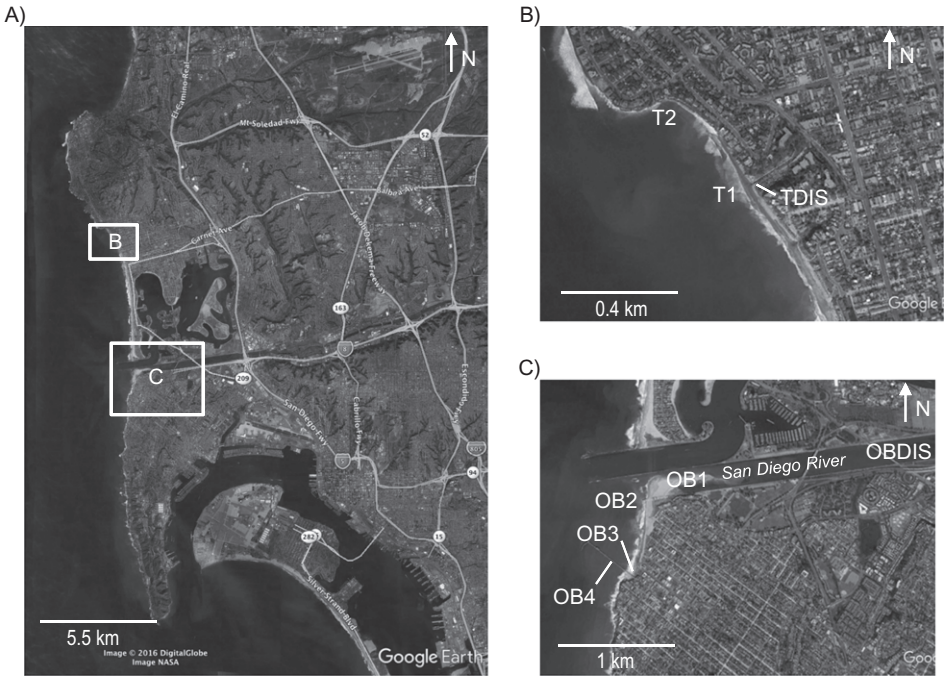


Figure 1. Monitoring beach water quality sampling locations in San Diego, California, winters of 2013–2014 and 2014–2015. Shown are the locations of the 2 monitored beaches along the San Diego coastline (A) and the water quality sampling sites at Tourmaline Surfing Park (B) and Ocean Beach (C). Samples were only collected at Ocean Beach and Tourmaline Surfing Park discharge locations (OBDIS and TDIS, respectively) during wet weather. Wet weather was defined as 0.25 cm or more of rain in 24 hours. T1 and T2, Tourmaline Surfing Park sampling sites 1 and 2; OB1–OB4, Ocean Beach sampling sites 1–4. Map Data: Google, Digital Globe, NASA.

Table 1. Characteristics of the Study Population by Mode of Enrollment, San Diego, California, 2013–2015

Characteristic	Beach[a]		Online[a]		Total	
	No.	%	No.	%	No.	%
No. of participants	89		565		654	
Participants with background survey	72	100	535	100	607	100
Age, years[b]						
18–30		35		35		35
31–40		22		26		26
41–50		11		16		16
≥51		29		13		15
Unreported		3		9		8
Female sex		19		21		21
College educated		68		63		63
Currently employed		74		76		75
Household income[b]						
<$15,000		11		6		7
$15,000–$35,000		15		10		11
$35,001–$50,000		11		7		7
$50,001–$75,000		8		13		12
$75,001–$100,000		17		14		14
$100,001–$150,000		17		14		14
>$150,000		7		13		12
Unreported		14		23		22
Days of surfing per week[b]						
≤1		11		15		14
2		12		18		17
3		26		26		26
4		26		20		21
≥5		24		18		19
Unreported		1		3		3
Chronic health conditions						
Ear problems		12		14		14
Sinus problems		7		8		8
Gastrointestinal condition		0		3		2
Respiratory condition		4		3		3
Skin condition		1		6		5
Allergies		10		16		15
Total days of observation	2,623	100	30,754	100	33,377	100
Days of observation by exposure						
Unexposed		46		47		47
Dry-weather exposure		48		43		43
Wet-weather exposure		6		10		10

[a] Beach enrollment only took place during the first winter (2013–2014); online enrollment spanned both winters (2013–2014 and 2014–2015). The study enrolled 73 individuals online during the first winter.

[b] Percentages within categories might not sum to 100 because of rounding.

to March 22, 2015; end of follow-up on April 16, 2015). We recruited surfers through postcards distributed at the monitored beaches and through an electronic newsletter distributed by the Surfrider Foundation's San Diego County chapter. Surfers were eligible if they were 18 years of age or older, could speak and read English, planned to surf in southern California

during the study period, had a valid e-mail address or mobile telephone number, and could access the internet with a computer or smartphone.

Participants completed a brief enrollment questionnaire, and each Tuesday they received a text message or e-mail reminder to complete a short weekly survey. Participants reported daily surf activity (location, date, and times of entry and exit) and illness symptoms (details below) for the previous 7 days using the study's web or smartphone (iOS or Android) application. We used an open cohort design in which participants were allowed to enter and exit the cohort over the follow-up period. We excluded follow-up time during which participants reported surfing outside of southern California. The study protocol was reviewed and approved by the institutional review board at the University of California, Berkeley, and all participants provided informed consent. Participants received a modest incentive for participation ($20 gift certificate per 4 weekly surveys completed). Web Table 1 (available at https://academic.oup.com/aje) includes a Strengthening the Reporting of Observational Studies in Epidemiology checklist.

Outcome definition and measurement

In weekly surveys, participants reported daily records of the following symptoms: diarrhea (defined as ≥3 loose/watery stools in 24 hours), sinus pain or infection, earache or infection, infection of an open wound, eye infection, skin rash, and fever. During the second winter, we added sore throat, cough, and runny nose. We created composite outcomes from the symptoms, including: gastrointestinal illness, which was defined as 1) diarrhea, 2) vomiting, 3) nausea and stomach cramps, 4) nausea and missed daily activities due to gastrointestinal illness, or 5) stomach cramps and missed daily activities due to gastrointestinal illness (15); and upper respiratory illness, which was defined as any 2 of the following: 1) sore throat, 2) cough, 3) runny nose, and 4) fever (16). We created a composite outcome of "any infectious symptom" defined as having any 1 of the following: gastrointestinal illness, diarrhea, vomiting, eye infection, infection of open wounds or fever. Our rationale was that it would exclude outcomes that could potentially have noninfectious causes (earache or infection, sinus pain or infection, skin rash, upper respiratory illness) and would capture a broad spectrum of sequelae associated with waterborne pathogens. We defined incident episodes as the onset of symptoms preceded by 6 or more symptom-free days to increase the likelihood that separate episodes represented distinct infections (17, 18).

Exposure definition and measurement

We classified the 3 days after each seawater exposure as exposed periods and all other days of observation as unexposed periods. We defined wet-weather exposure as exposure to seawater within 3 days of 0.25 cm or more of rainfall in a 24-hour period, which is the rainfall criterion used by San Diego County for posting wet-weather beach advisories; we classified all other seawater exposure as dry-weather exposure. We used rainfall measurements from the National Oceanic and Atmospheric Administration Lindbergh Field Station. Among surfers, most exposure took place during the morning hours, so if a storm's precipitation started after 12:00 PM, we did not classify that day as wet weather (only the following day) to reduce exposure misclassification.

Staff collected daily water samples from January 15, 2014, to March 5, 2014, and from December 2, 2014, to March 31, 2015, at 6 sites across the 2 monitored beaches (Figure 1). Staff collected 1-liter water samples in the morning (08:30 AM ± 2 hours) just below the water surface (0.5–1.0 meters) in sterilized, sample-rinsed bottles. We sampled discharges during 6 rainstorms immediately upstream from where Tourmaline Creek and the San Diego River discharge to the sea (Figure 1). We tested samples for culturable *Enterococcus* (US Environmental Protection Agency method 1600), fecal coliforms (standard method 9222D), and total coliforms (standard method 9222B). All laboratory analyses met quality-control objectives for absence of background contamination (blanks) and precision (duplicates).

Statistical analysis

We prespecified all analyses (19). Web Appendices 1 and 2 contain statistical details and sample size calculations. In the seawater exposure analysis, we calculated incidence rates by dividing incident episodes by person-days in unexposed and exposed periods during follow-up. If participants missed weekly surveys during follow-up, we did not include those periods in the analysis. We measured the association between seawater exposure and subsequent illness using an incidence rate ratio, which we estimated using a log-linear rate model with robust standard errors to account for repeated observations within individuals (20, 21). To examine illness rates separately for dry- and wet-weather exposures, we created a 3-level categorical exposure that classified each participant's follow-up time into unexposed, dry-weather exposure, and wet-weather exposure periods. We calculated a log-linear test of trend in the incidence rate ratios for dry- and wet-weather exposures (22).

In the fecal indicator association analysis, we estimated the association between levels of fecal indicator bacteria and illness using the subset of surf sessions matched to water-quality indicator measurements at the monitored beaches. We matched daily geometric mean indicator levels to surfers by beach and date (weighted by time in water if recent exposure included multiple days).We modeled the relationship between indicator levels and illness using a log-linear model and estimated the incidence rate ratio associated with a $1-\log_{10}$ increase in indicator level. We also estimated the incidence rate ratio associated with exposures to water above versus below US Environmental Protection Agency regulatory guidelines (geometric mean *Enterococcus* >35 colony-forming units per 100 mL) (23) or, in a second definition, if any single sample on the exposure day exceeded 104 colony-forming units per 100 mL. We hypothesized that the relationship between fecal indicator bacteria and illness could be modified by dry- or wet-weather exposure and allowed the exposure-response relationship to vary during dry and wet weather by including an indicator for wet-weather periods and a term for the interaction between indicator bacteria levels and the indicator of wet weather. We controlled for potential confounding (24) from demographic,

exposure-related, and baseline health characteristics (Web Appendix 1). In Web Appendices 3–6 we describe additional analyses, including conversion of estimates to the absolute risk scale, sensitivity analyses, and negative control exposure analyses (25, 26).

RESULTS

Study population

We enrolled 654 individuals who contributed on average 51 days of follow-up (range, 6–139 days). The study population's median age was 34 years (interquartile range, 27–45), and the majority of participants were male (73%), college-educated (63%), and employed (75%) (Table 1). Follow-up included 33,377 person-days of observation after excluding time spent outside of southern California (623 person-days). We excluded from adjusted analyses 47 individuals (1,599 person-days of observation) who provided outcome and exposure information but failed to complete a background questionnaire and thus had missing covariate information.

Water quality and surfer exposure

There were 10 rainstorms with 0.25 cm or more of rain during the study. Field staff collected 1,073 beach water samples and 92 wet-weather discharge samples for fecal indicator bacteria analysis. Median *Enterococcus* levels were higher during wet weather than during dry weather (Figure 2). During follow-up, surfers entered the ocean twice per week on average and experienced 10,081 total days of seawater exposure, including 1,327 days of wet-weather exposure. Surfers were less likely to enter the ocean during or within 1 day of rain. The median ocean entry time was 08:00 AM (interquartile range, 06:45–10:30 AM), and the median time spent in the water was 2 hours (interquartile range, 1–2 hours) (Web Figure 1). Of the 10,081 exposure days, surfers reported wearing a wetsuit during 95%, immersing their head during 96%, and swallowing water during 38%. The most frequented surf locations were the 2 monitored beaches: Tourmaline Surfing Park (25% of surf days) and Ocean Beach (16% of surf days), which reflected targeted enrollment at those beaches (Web Figure 2). There were 5,819 days of observation matched to water-quality measurements at monitored beaches, including 1,358 days during wet weather.

Illness associated with seawater exposure

Seawater exposure in the past 3 days was associated with increased incidence rates of all outcomes except for upper respiratory illness (Web Table 2). Unadjusted and adjusted incidence rate ratio estimates were similar, and for most outcomes, adjusted incidence rate ratios were slightly attenuated toward the null (Web Table 2). With the exception of fever and skin rash, incidence rates increased from unexposed to dry-weather exposure to wet-weather exposure periods (Table 2), a pattern also present on the risk scale (Web Figure 3). Compared with unexposed periods, wet-weather exposure led to the largest relative increase in earaches/infections (Table 3; adjusted incidence rate ratio (IRR) = 3.28, 95% confidence interval (CI): 1.95, 5.51) and infection of open wounds (Table 3; adjusted IRR: 4.96, 95% CI: 2.18, 11.29). Sensitivity analyses that shortened the wet-weather window increased the difference between dry- and wet-weather incidence rates for most outcomes (Web Figure 4).

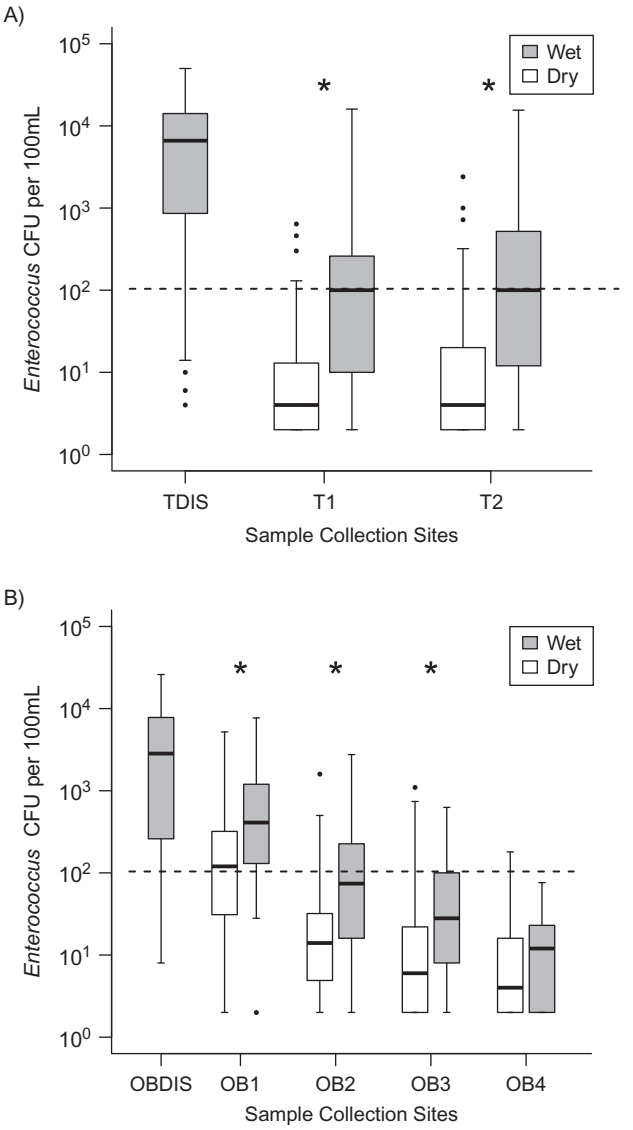


Figure 2. *Enterococcus* levels during dry and wet weather at the sampling locations at Tourmaline Surfing Park (A) and Ocean Beach (B) mapped in Figure 1. Boxes mark interquartile ranges, vertical lines mark 1.5 times the interquartile range, and points mark outliers. Horizontal dashed lines mark the single-sample California recreational water quality guideline (104 CFU/100 mL). Asterisks (*) identify sampling locations with levels that differ between wet and dry periods based on a 2-sample, 2-sided t-test ($P < 0.05$) assuming unequal variances. Samples were only collected at Ocean Beach and Tourmaline Surfing Park discharge locations (OBDIS and TDIS, respectively) during wet weather. Wet weather was defined as 0.25 cm or more of rain in 24 hours. CFU, colony-forming units; T1 and T2, Tourmaline Surfing Park sampling sites 1 and 2; OB1–OB4, Ocean Beach sampling sites 1–4.

Table 2. Incidence Rates Among Surfers by Type of Seawater Exposure, San Diego, California, 2013–2015

Outcome	Unexposed Periods			Dry-Weather Exposure			Wet-Weather Exposure[a]		
	No. of Episodes	No. of Days at Risk	Rate per 1,000	No. of Episodes	No. of Days at Risk	Rate per 1,000	No. of Episodes	No. of Days at Risk	Rate per 1,000
Gastrointestinal illness	90	14,884	6.0	116	13,769	8.4	31	3,037	10.2
Diarrhea	75	15,086	5.0	88	13,909	6.3	27	3,061	8.8
Sinus pain or infection	109	14,475	7.5	139	13,391	10.4	37	2,998	12.3
Earache or infection	59	14,931	4.0	111	13,618	8.2	37	3,008	12.3
Infection of open wound	14	15,456	0.9	30	14,080	2.1	11	3,119	3.5
Skin rash	42	15,024	2.8	66	13,750	4.8	15	3,007	5.0
Fever	51	15,156	3.4	69	14,138	4.9	6	3,152	1.9
Upper respiratory illness[b]	117	12,001	9.7	111	11,025	10.1	31	2,543	12.2
Any infectious symptom[c]	138	14,445	9.6	181	13,176	13.7	47	2,926	16.1

[a] Defined as entering the sea within 3 days of 0.25 cm or more of rain in 24 hours.
[b] Only measured in year 2 of the study.
[c] Includes gastrointestinal illness, eye infections, infected wounds, and fever.

Illness associated with fecal indicator bacteria levels

Enterococcus, total coliform, and fecal coliform levels were positively associated with increased incidence of almost all outcomes during the study (Web Table 3). Rainfall was a strong effect modifier of the association (Table 4). During dry weather, there was no association between *Enterococcus* levels and illness except for infected wounds, but *Enterococcus* was strongly associated with illness after wet-weather exposure (e.g., for each $\log_{10}$ increase, gastrointestinal illness IRR = 2.17, 95% CI: 1.16, 4.03; Table 4, Web Figure 5, and Web Table 4). Associations were attenuated in adjusted analyses, but relationships were similar (e.g., for gastrointestinal illness, wet-weather IRR = 1.75, 95% CI: 0.80, 3.84; Table 4). There was evidence for excess risk of gastrointestinal illness at higher *Enterococcus* levels only during wet-weather periods (Web Figure 6): The predicted excess risk that corresponded to the current US Environmental Protection Agency regulatory guideline of 35 colony-forming units per 100 mL was 16 episodes per 1,000 (95% CI: 5, 27). Negative control analyses showed no consistent association between fecal indicator bacteria and illness among participants during periods in which they had no recent seawater contact (Web Table 5).

DISCUSSION

Key results

To our knowledge, this is the first prospective cohort study in which the association between incident illness and exposure to seawater in wet weather has been measured, and the findings represent novel empirical measures of incident illness associated with storm water discharges. There was a consistent increase in acute illness incidence rates between unexposed, dry-weather, and wet-weather exposure periods (Tables 2 and 3). Rainstorms led to higher levels of fecal indicator bacteria (Figure 2), and a sensitivity analysis illustrated that a 2–3 day window after rainstorms captured the majority of excess incidence associated with wet-weather exposure (Web Figure 4). Fecal indicator bacteria matched to individual surf sessions were strongly associated with illness only during wet weather periods (Table 4, Web Figure 5).

Interpretation

Swimmers are more rare during the winter months, and surfers' frequent and intense exposure made them an ideal population in which to study the relationship between illness and exposure to seawater in wet weather (27). The associations estimated in this study may not reflect those of the general population, but among a highly exposed subgroup of athletes, our results measure the illness associated with seawater exposure after rainstorms in southern California. Enrolling surfers led to some important differences between the present study population and most swimmer cohorts. We enrolled adults because we could not guarantee adequate consent for minors through online enrollment, whereas swimmer cohorts have historically enrolled predominantly families with children (28); children are more susceptible and have greater risk than do adult swimmers (15). Participants surfed twice per week for 2 hours each session, with nearly universal head immersion (96% of exposures) and frequent water ingestion (38% of exposures). This far exceeds exposure levels recorded in swimmer cohorts. Likely because of surfers' repeated exposures to pathogens in seawater, studies have found higher levels of immunity to hepatitis A and more frequent gut colonization by antibiotic-resistant *Escherichia coli* among surfers than among the general population (29, 30).

Despite surfers' intense and frequent exposures, gastrointestinal illness rates observed in the present study were similar to those measured among beachgoers California cohorts in the summer (Web Appendix 6, Web Figure 7), and the increase in gastrointestinal illness rates associated with seawater exposure (adjusted IRR = 1.33, 95% CI: 0.99, 1.78; Web Table 2) was similar to estimates measured in marine swimmer cohorts in California and elsewhere in the United States (15, 31). However, the 3-fold increase in rates of

Table 3. Incidence Rate Ratios for Surfer Illnesses Within 3 Days of Dry- and Wet-Weather Seawater Exposure Compared With Unexposed Periods, San Diego, California, 2013–2015

Outcome	Unadjusted[a]				Adjusted[a,b]			
	Dry Weather		Wet Weather[c]		Dry Weather		Wet Weather[c]	
	IRR	95% CI	IRR	95% CI	IRR	95% CI	IRR	95% CI
Gastrointestinal illness	1.39	1.05, 1.86	1.69	1.10, 2.59	1.30	0.95, 1.76	1.41	0.92, 2.17
Diarrhea	1.27	0.92, 1.76	1.77	1.11, 2.83	1.22	0.86, 1.73	1.51	0.95, 2.41
Sinus pain or infection	1.38	1.05, 1.80	1.64	1.12, 2.40	1.23	0.93, 1.64	1.51	1.01, 2.26
Earache or infection	2.06	1.47, 2.90	3.11	1.94, 4.98	1.86	1.27, 2.71	3.28	1.95, 5.51
Infection of open wound	2.35	1.27, 4.36	3.89	1.83, 8.30	3.04	1.54, 5.98	4.96	2.18, 11.29
Skin rash	1.72	1.16, 2.54	1.78	0.98, 3.24	1.64	1.11, 2.41	1.80	0.97, 3.35
Fever	1.45	0.99, 2.12	0.57	0.24, 1.31	1.56	1.04, 2.34	0.64	0.27, 1.52
Upper respiratory illness[d]	1.03	0.79, 1.35	1.25	0.84, 1.86	1.04	0.79, 1.36	1.17	0.79, 1.74
Any infectious symptom[e]	1.44	1.14, 1.82	1.68	1.19, 2.38	1.50	1.17, 1.92	1.62	1.14, 2.30

Abbreviations: CI, confidence interval; IRR, incidence rate ratio.

[a] Unadjusted and adjusted incidence rate ratios compare incidence rates in the 3 days after seawater exposure during dry or wet weather with incidence rates during unexposed periods. Table 2 includes the underlying data. Tests of trend in the IRR between exposure categories are significant ($P < 0.05$) if the confidence interval for wet-weather exposure excludes 1.0 (22).

[b] We controlled for the following time-invariant potential confounders: age, sex, educational level, employment status, household income, years the individual had surfed, reported behavior of typically avoiding the ocean after wet weather, surfboard length, mode of enrollment (beach vs. Internet). We controlled for chronic health conditions only for the corresponding outcomes: ear problems, sinus problems, gastrointestinal conditions, respiratory conditions, skin conditions. We also controlled for the following time-varying potential confounders: entered the ocean for an activity other than surfing, any illness symptoms in the week preceding the risk window, day of recall, day of the week, and rainfall total during the past 3 days.

[c] Defined as entering the sea within 3 days of 0.25 cm or more of rain in 24 hours.

[d] Only measured in year 2 of the study.

[e] Includes gastrointestinal illness, eye infections, infected wounds, and fever.

earache/infection and 5-fold increase in infected open wounds associated with exposure after rainstorms (Table 3) are stronger associations than have been reported in previous studies, and they provide evidence for increased incidence of a broad set of infectious symptoms after seawater exposure within 3 days of rain.

Fecal indicator bacteria were a reliable marker of human illness risk in this setting only within 3 days of rainfall (Table 4). Our results are consistent with summer studies in California in which investigators found associations between *Enterococcus* levels and illness only if there was a well-defined source of human fecal contamination (4–8). Our findings are also consistent with model predictions of higher gastrointestinal illness risk among southern California surfers after storms (32). Molecular testing for pathogens in storm water discharge to study monitored beaches identified near-ubiquitous presence of norovirus and *Campylobacter* species, and models parameterized with pathogen measurements predicted higher illness risk after rainstorms (14). The association between fecal indicator bacteria measured during wet weather and a range of nonenteric illnesses, such as sinus pain or infection and fever (Table 4), suggests that fecal indicator bacteria may mark broader bacterial or viral pathogen contamination in seawater after rainstorms.

Some study outcomes could have noninfectious causes associated with surfing. Earache and sinus pain can result from physical incursion of saltwater through surfing's high-intensity exposure, ingestion of saltwater can cause gastrointestinal symptoms, and wetsuit use could cause skin rashes. If the association between surf exposure and symptoms resulted from noninfectious causes, we would expect similar incidence rates after wet- and dry-weather exposures. This was observed for skin rash, but incidence rates for sinus, ear, and gastrointestinal illnesses were higher after wet-weather exposure (Table 2), and the strong association between fecal indicator bacteria and fever during wet-weather conditions was consistent with an infectious etiology (Table 4).

It is also possible that some infections acquired during surfing could result from nonanthropogenic sources. The ocean was warmer than usual during the second winter because of a weak El Niño, which caused conditions favorable to naturally occurring *Vibrio parahaemolyticus* and toxin-producing marine algae that can cause human illness (33). Wound infection was the single outcome strongly associated with fecal indicator bacteria measured during dry weather (Table 4), an observation consistent with a pathogen source like *V. parahaemolyticus* that covaries with fecal indicator bacteria even in nonstorm conditions. Yet, the consistently higher rates of infected wounds and other symptoms after wet-weather exposure compared with dry-weather exposure (Tables 2 and 3) suggests that storm water runoff impacted by anthropogenic sources constitutes an important pathogen source in this setting.

Table 4. Surfer Illness Associated With a $\log_{10}$ Increase in Fecal Indicator Bacteria Levels, Stratified by Exposure During Dry and Wet Weather, Tourmaline Surfi
San Diego, California, 2013–2015

Fecal Indicator Bacteria and Illness Symptom	Dry Weather		Wet Weather		Unadjusted					Ad		
					Dry Weather		Wet Weather		*P* Value[b]	Dry Weather		
	Episodes	Days at Risk	Episodes	Days at Risk	IRR	95% CI	IRR	95% CI		IRR	95% CI	I
Enterococcus												
Gastrointestinal illness	30	4,251	10	1,297	0.86	0.47, 1.58	2.17	1.16, 4.03	0.04	0.85	0.46, 1.56	1.
Diarrhea	24	4,285	9	1,305	1.13	0.62, 2.07	2.38	1.27, 4.46	0.11	1.16	0.63, 2.14	2.
Sinus pain or infection	44	4,130	19	1,262	1.34	0.79, 2.26	1.93	1.17, 3.19	0.33	0.96	0.53, 1.76	1.
Earache or infection	38	4,233	14	1,274	0.74	0.37, 1.47	1.23	0.50, 3.02	0.38	0.70	0.35, 1.40	1.
Infection of open wound	19	4,360	6	1,332	2.69	1.05, 6.90	2.24	0.65, 7.69	0.83	2.79	1.12, 6.95	2.
Skin rash	19	4,230	5	1,267	1.46	0.68, 3.14	0.89	0.21, 3.82	0.56	1.09	0.42, 2.80	0.
Fever	22	4,366	2	1,342	1.33	0.69, 2.56	3.29	2.35, 4.59	0.01	1.29	0.66, 2.52	3.
Upper respiratory illness[c]	37	3,679	15	1,090	0.89	0.55, 1.45	1.94	0.85, 4.42	0.10	0.74	0.44, 1.25	1.
Any infectious symptom[d]	50	4,080	17	1,264	1.12	0.69, 1.83	2.51	1.49, 4.24	0.04	1.06	0.64, 1.76	2.
Fecal coliforms												
Gastrointestinal illness	30	4,251	10	1,297	0.82	0.42, 1.61	2.96	1.50, 5.83	0.01	0.76	0.38, 1.54	2.
Diarrhea	24	4,285	9	1,305	1.04	0.53, 2.04	3.34	1.72, 6.47	0.02	1.05	0.51, 2.16	3.
Sinus pain or infection	44	4,130	19	1,262	1.57	0.87, 2.84	2.18	1.11, 4.26	0.48	0.75	0.35, 1.58	1.
Earache or infection	38	4,233	14	1,274	0.83	0.39, 1.76	1.46	0.63, 3.39	0.29	0.99	0.51, 1.92	1.
Infection of open wound	19	4,360	6	1,332	2.76	0.91, 8.36	2.67	0.85, 8.41	0.97	3.21	1.03, 10.03	4.
Skin rash	19	4,230	5	1,267	1.69	0.72, 3.99	1.03	0.24, 4.43	0.56	1.18	0.39, 3.56	0.
Fever	22	4,366	2	1,342	1.15	0.49, 2.70	4.99	3.19, 7.79	0.00	1.16	0.49, 2.73	6.
Upper respiratory illness[c]	37	3,679	15	1,090	0.97	0.50, 1.89	2.33	0.75, 7.23	0.19	0.73	0.38, 1.40	2.
Any infectious symptom[d]	50	4,080	17	1,264	1.17	0.69, 1.97	3.21	1.84, 5.58	0.01	1.11	0.65, 1.91	3.
Total coliforms												
Gastrointestinal illness	30	4,251	10	1,297	0.77	0.40, 1.47	2.62	1.63, 4.24	0.01	0.83	0.42, 1.63	1.
Diarrhea	24	4,285	9	1,305	0.66	0.29, 1.51	2.59	1.53, 4.38	0.02	0.78	0.35, 1.70	1.
Sinus pain or infection	44	4,130	19	1,262	1.52	0.84, 2.77	2.02	1.04, 3.93	0.55	1.08	0.54, 2.19	1.
Earache or infection	38	4,233	14	1,274	1.03	0.54, 1.96	1.67	0.63, 4.41	0.40	0.92	0.46, 1.82	1.
Infection of open wound	19	4,360	6	1,332	3.46	0.79,15.20	2.16	0.46, 10.16	0.69	4.02	0.91, 17.67	2.
Skin rash	19	4,230	5	1,267	1.58	0.73, 3.40	1.14	0.34, 3.81	0.65	1.30	0.48, 3.53	1.
Fever	22	4,366	2	1,342	1.59	0.78, 3.22	7.48	4.28, 13.08	0.00	1.62	0.77, 3.37	9.
Upper respiratory illness[a]	37	3,679	15	1,090	0.87	0.49, 1.52	2.04	0.84, 4.96	0.12	0.72	0.40, 1.30	1.
Any infectious symptom[d]	50	4,080	17	1,264	1.35	0.78, 2.34	3.26	1.76, 6.01	0.06	0.69	0.23, 2.07	3.

Abbreviations: CI, confidence interval; IRR, incidence rate ratio.

[a] We controlled for the following time-invariant potential confounders: age, sex, educational level, employment status, household income, years the individual had surfed, reported ocean after wet weather, surfboard length, mode of enrollment (beach vs. Internet). We controlled for chronic health conditions only for the corresponding outcomes: ear problems, sinus tions, respiratory conditions, skin conditions. We also controlled for the following time-varying potential confounders: entered the ocean for an activity other than surfing, any illness sy risk window, day of recall, day of the week, and rainfall total during the past 3 days.

[b] *P* value for multiplicative effect modification of dry versus wet weather.

[c] Only measured in year 2 of the study.

[d] Includes gastrointestinal illness, eye infections, infected wounds, and fever.

ing Park and Ocean Beach,

djusted[a]			
Wet Weather			
RR	95% CI	P Value[b]	
.75	0.80, 3.84	0.16	
.00	0.92, 4.82	0.08	
.61	0.96, 2.69	0.07	
.32	0.51, 3.41	0.22	
.94	0.79, 10.97	0.31	
.51	0.06, 4.04	0.95	
.53	2.37, 5.24	0.50	
.89	0.87, 4.11	0.01	
.52	1.41, 4.50	0.06	
.59	1.02, 6.56	0.03	
.20	1.31, 7.85	0.04	
.52	0.62, 3.73	0.08	
.59	0.84, 3.01	0.22	
.12	0.95, 17.91	0.92	
.54	0.09, 3.06	0.79	
.22	3.88, 9.96	0.42	
.03	0.70, 5.89	0.00	
.42	1.76, 6.66	0.11	
.96	1.22, 3.15	0.01	
.99	1.19, 3.35	0.08	
.79	0.93, 3.42	0.09	
.72	0.64, 4.61	0.28	
.38	0.60, 9.45	0.32	
.11	0.28, 4.41	0.03	
.24	4.64, 18.41	0.86	
.87	0.84, 4.19	0.00	
.02	1.56, 5.38	0.08	

l behavior of typically avoiding the
s problems, gastrointestinal condi-
mptoms in the week preceding the

Limitations

The use of self-reported symptoms could bias the association between seawater exposure and illness away from the null if surfers overreported illness after exposure; conversely, random (nondifferential) errors in exposures or outcomes could bias associations toward the null (34). The survey measured daily exposure and outcomes in separate modules—an intentional decision to separate the measurements and inhibit systematic reporting bias. Adjusted analyses controlled for day of recall and day of the week to reduce nondifferential bias from recall errors but would not control for systematic bias. Negative control exposure analyses found no association between *Enterococcus* levels and illness on days with no recent water exposure (Web Table 5), which suggests that unmeasured confounding or reporting bias is unlikely to explain the association between *Enterococcus* levels and illness. Moreover, the use of daily average levels of fecal indicator bacteria could bias the association between water quality and illness toward the null if the averaging resulted in nondifferential misclassification error (35).

We measured incident outcomes within 3 days of seawater exposure because the population regularly entered the ocean, a 3-day period captures the incubation period for the most common waterborne pathogens (e.g., norovirus, *Campylobacter* species, *Salmonella* species) (36), and past studies found that most excess episodes of gastrointestinal illness associated with seawater exposure occurred in the first 1–2 days (15). Illness caused by waterborne pathogens with longer incubation periods (e.g., *Cryptosporidium* species) (37) could have been misclassified in this study, which could bias results toward the null by artificially increasing incidence rates in unexposed periods and decreasing rates in exposed periods.

Conclusions

Surfing was associated with increased incidence of several categories of symptoms, and associations were stronger if surfing took place shortly after rainstorms. Higher levels of fecal indicator bacteria were strongly associated with fever, sinus pain/infection, wound infection, and gastrointestinal symptoms within 3 days of rainstorms. The internal consistency between water quality measurements, patterns of illness after dry- and wet-weather exposures, and incidence profiles with time since rainstorms lead us to conclude that seawater exposure during or close to rainstorms at beaches impacted by urban runoff in southern California increases the incidence rates of a broad set of acute illnesses among surfers. These findings provide strong evidence to support the posting of beach warnings after rainstorms and initiatives that would reduce pathogen sources in urban runoff that flows to coastal waters.

ACKNOWLEDGMENTS

Author affiliations: Division of Epidemiology, School of Public Health, University of California, Berkeley, California (Benjamin F. Arnold, Ayse Ercumen, Jade Benjamin-Chung, John M. Colford, Jr.); Southern California Coastal Water Research Project, Costa Mesa, California (Kenneth C. Schiff, Joshua A. Steele, John F. Griffith, Steven J. Steinberg, Paul Smith, Stephen B. Weisberg); Orange County Sanitation District, Fountain Valley, California (Charles D. McGee; retired); and Surfrider Foundation, San Clemente, California (Richard Wilson, Chad Nelsen).

The study was funded by the city and county of San Diego, California.

We thank the field team members who enrolled participants at the beach and collected water samples throughout the study. We also thank Laila Othman, Sonji Romero, Aaron Russell, Joseph Toctocan, Laralyn Asato, Zaira Valdez, and the staff at City of San Diego Marine Microbiology Laboratory who generously provided laboratory space to test water specimens, and Jeffrey Soller, Mary Schoen, and members of the study's external advisory committee for earlier comments on the results.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of interest: none declared.

REFERENCES

1. Noble RT, Weisberg SB, Leecaster MK, et al. Storm effects on regional beach water quality along the southern California shoreline. *J Water Health*. 2003;1(1):23–31.
2. Leecaster MK, Weisberg SB. Effect of sampling frequency on shoreline microbiology assessments. *Mar Pollut Bull*. 2001; 42(11):1150–1154.
3. Ackerman D, Weisberg SB. Relationship between rainfall and beach bacterial concentrations on Santa Monica bay beaches. *J Water Health*. 2003;1(2):85–89.
4. Haile RW, Witte JS, Gold M, et al. The health effects of swimming in ocean water contaminated by storm drain runoff. *Epidemiology*. 1999;10(4):355–363.
5. Colford JM, Wade TJ, Schiff KC, et al. Water quality indicators and the risk of illness at beaches with nonpoint sources of fecal contamination. *Epidemiology*. 2007;18(1): 27–35.
6. Colford JM, Schiff KC, Griffith JF, et al. Using rapid indicators for Enterococcus to assess the risk of illness after exposure to urban runoff contaminated marine water. *Water Res*. 2012; 46(7):2176–2186.
7. Arnold BF, Schiff KC, Griffith JF, et al. Swimmer illness associated with marine water exposure and water quality indicators: impact of widely used assumptions. *Epidemiology*. 2013;24(6):845–853.
8. Yau VM, Schiff KC, Arnold BF, et al. Effect of submarine groundwater discharge on bacterial indicators and swimmer health at Avalon Beach, CA, USA. *Water Res*. 2014;59:23–36.
9. Taylor K. Contagion Present. Surfer Magazine. http://www.surfermag.com/features/contagion-present. Published July 20, 2016. Accessed August 17, 2016.
10. Dwight RH, Baker DB, Semenza JC, et al. Health effects associated with recreational coastal water use: urban versus rural California. *Am J Public Health*. 2004;94(4):565–567.
11. Harding AK, Stone DL, Cardenas A, et al. Risk behaviors and self-reported illnesses among Pacific Northwest surfers. *J Water Health*. 2015;13(1):230–242.

12. Stormsurf. Weather basics. http://www.stormsurf.com/page2/tutorials/weatherbasics.shtml. Published September 26, 2003. Accessed October 27, 2016.
13. Heal the Bay. Heal the Bay's 2014-2015 Annual Beach Report Card. Santa Monica, CA: Heal the Bay; 2015. http://www.healthebay.org/sites/default/files/BRC_2015_final.pdf. Accessed December 5, 2016.
14. Schiff K, Griffith J, Steele J, et al. The Surfer Health Study: A Three-Year Study Examining Illness Rates Associated With Surfing During Wet Weather. Costa Mesa, CA: Southern California Coastal Water Research Project; 2016. http://ftp.sccwrp.org/pub/download/DOCUMENTS/TechnicalReports/943_SurferHealthStudy.pdf. Published September 20, 2016. Accessed December 5, 2016.
15. Arnold BF, Wade TJ, Benjamin-Chung J, et al. Acute gastroenteritis and recreational water: highest burden among young US children. *Am J Public Health*. 2016;106(9):1690–1697.
16. Wade TJ, Sams E, Brenner KP, et al. Rapidly measured indicators of recreational water quality and swimming-associated illness at marine beaches: a prospective cohort study. *Environ Health*. 2010;9:66.
17. Colford JM, Wade TJ, Sandhu SK, et al. A randomized, controlled trial of in-home drinking water intervention to reduce gastrointestinal illness. *Am J Epidemiol*. 2005;161(5):472–482.
18. Colford JM, Hilton JF, Wright CC, et al. The Sonoma Water Evaluation Trial: a randomized drinking water intervention trial to reduce gastrointestinal illness in older adults. *Am J Public Health*. 2009;99(11):1988–1995.
19. Arnold B, Ercumen A. The Surfer Health Study. Open Science Framework. https://osf.io/hvn7s. Published July 29, 2015. Updated July 29, 2016. Accessed December 5, 2016.
20. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott Williams and Wilkins; 2008.
21. Zou G. A modified poisson regression approach to prospective studies with binary data. *Am J Epidemiol*. 2004;159(7):702–706.
22. Vittinghoff E, Glidden DV, Shiboski SC, et al. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. 2nd ed. New York, NY: Springer Science & Business Media; 2012.
23. United States Environmental Protection Agency. *Recreational Water Quality Criteria*. Washington, DC: United States Environmental Protection Agency; 2012. (Office of Water publication no. 820-F-12-058). https://www.epa.gov/sites/production/files/2015-10/documents/rwqc2012.pdf. Accessed January 24, 2017.
24. VanderWeele TJ, Shpitser I. A new criterion for confounder selection. *Biometrics*. 2011;67(4):1406–1413.
25. Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*. 2010;21(3):383–388.
26. Arnold BF, Ercumen A, Benjamin-Chung J, et al. Brief report: negative controls to detect selection bias and measurement bias in epidemiologic studies. *Epidemiology*. 2016;27(5):637–641.
27. Rothman KJ, Gallacher JE, Hatch EE. Why representativeness should be avoided. *Int J Epidemiol*. 2013;42(4):1012–1014.
28. Wade TJ, Pai N, Eisenberg JN, et al. Do US Environmental Protection Agency water quality guidelines for recreational waters prevent gastrointestinal illness? A systematic review and meta-analysis. *Environ Health Perspect*. 2003;111(8):1102–1109.
29. Gammie A, Morris R, Wyn-Jones AP. Antibodies in crevicular fluid: an epidemiological tool for investigation of waterborne disease. *Epidemiol Infect*. 2002;128(2):245–249.
30. Leonard A. *Are Bacteria in the Coastal Zone a Threat to Human Health?* [dissertation]. Exeter, UK: University of Exeter; 2016. https://ore.exeter.ac.uk/repository/handle/10871/22805. Accessed October 14, 2016.
31. Fleisher JM, Fleming LE, Solo-Gabriele HM, et al. The BEACHES Study: health effects and exposures from non-point source microbial contaminants in subtropical recreational marine waters. *Int J Epidemiol*. 2010;39(5):1291–1298.
32. Tseng LY, Jiang SC. Comparison of recreational health risks associated with surfing and swimming in dry weather and post-storm conditions at Southern California beaches using quantitative microbial risk assessment (QMRA). *Mar Pollut Bull*. 2012;64(5):912–918.
33. Van Dolah FM. Marine algal toxins: origins, health effects, and their increased occurrence. *Environ Health Perspect*. 2000;108(suppl 1):133–141.
34. Copeland KT, Checkoway H, McMichael AJ, et al. Bias due to misclassification in the estimation of relative risk. *Am J Epidemiol*. 1977;105(5):488–495.
35. Fleisher JM. The effects of measurement error on previously reported mathematical relationships between indicator organism density and swimming-associated illness: a quantitative estimate of the resulting bias. *Int J Epidemiol*. 1990;19(4):1100–1106.
36. Widdowson MA, Sulka A, Bulens SN, et al. Norovirus and foodborne disease, United States, 1991-2000. *Emerg Infect Dis*. 2005;11(1):95.
37. Jokipii L, Jokipii AM. Timing of symptoms and oocyst excretion in human cryptosporidiosis. *N Engl J Med*. 1986;315(26):1643–1647.