ISES Europe Training Series

DoE 9: Statistics and Epidemiology

Module 3: Introduction to Epidemiology

Hello and welcome to this video lecture, which is part of Domain of Expertise 9, Statistics and Epidemiology, in the ISES Europe training series.

Just a brief legal notice before we start: you need to have permission to reuse these materials, and we retain the copyright.

So just a reminder that this is the final of nine domains of expertise, and you can access the other videos on the ISES Europe website.

So, Domain of Expertise 9 includes three modules: one on an introduction to statistics, one on statistics for exposure scientists, and Module 3, the one you're watching now, is a brief introduction to epidemiology.

Just a reminder that this is an introduction to quite a big topic, so if you want to learn more about it, I encourage you to.

My name is Ruairí Weiner. I'm a Teaching Fellow in Research Methods and Statistics at the University College Dublin School of Public Health, Physiotherapy, and Sports Science.

I'd just encourage you to read this brief disclaimer. It's telling you that this lecture is an introductory framework, squeezed into a short video, so topics have to be simplified for ease of understanding. We try to make sure everything is accurate, but of course errors can occur, and this doesn't represent any organization's official position.

So the learning objectives for today are to understand an overview of the history and aims of epidemiology, to describe the basics of measuring exposures and risk, and to understand the role of epidemiology in linking environmental exposures to negative health outcomes.

So, we have five topics and a summary. Our first topic is defining epidemiology.

So the word itself comes from Greek. I'm not an expert on Greek, but I understand that *epi* means on or upon, and *demos* means people, and *logos* means study. So essentially, we're talking about the study of the population, its condition, what's going on for the population.

And I quite like the US CDC definition, which is that we study the distribution of health events and the determinants of health.

So the distribution: we count health events, things like diagnoses, deaths, accidents, and we map their distribution in the population.

So: what locations do these events occur in? What age groups, gender, socioeconomic status do we see conditions or deaths, etc.? And are rates of a health condition or event relatively higher or lower in some groups? Are there some age groups or genders who are diagnosed more often with the condition?

And then determinants. This is about mapping health events to reveal risk factors. So those are groups who are more at risk, or exposures that raise the risk of a health event. Then we ask, can we intervene on any risk factor to reduce rates? Can I target an intervention at an at-risk group to protect them from a negative outcome?

So, to do that, we need to identify the determinants of health that can be targeted with an intervention like a policy, a treatment, an education campaign. So determinants are really causes of diseases: things that, if I turn them off or turn them down, that changes the outcome.

So now we look at the foundations of epidemiology.

And when I'm explaining epidemiology, I like to go back to our founding figure of John Snow, because during his career, he kind of exemplified what are still the key techniques in epidemiology in his work.

So, who was John Snow? He was a Victorian obstetrician, and he had a habit of noticing things that appeared to make people sick and devising ways to prove his suspicions. And through that work, he ended up setting up a lot of the core methods in epidemiology through his efforts to devise ways of proving what might be making people sick.

So he found ways of tracking and mapping disease outbreaks, and identifying explanations for those outbreaks.

So early in his career in medical school, he did a bit of innovation around experiments to demonstrate health risks for humans. While studying, he noticed a lot of medical students had gastrointestinal complaints while they were conducting autopsies.

And recently, there had been a switch to embalming using arsenic, even though arsenic was a known toxin at the time.

So this raised Snow's suspicions, so he conducted experiments that showed that arsenic gas is released from the embalmed bodies.

And his clever experiment led to a policy change. The medical school stopped using arsenic in embalming, and the sale of arsenic candles was stopped. If you can imagine, there were arsenic candles; I think they burned brighter or something.

So that's an early example of someone using observation and experimentation to demonstrate the danger of an environmental exposure, leading to an effective intervention or policy change, even though I think they might have switched back to arsenic later.

Then cholera was what John Snow was most famous for. So during his career, cholera was a major problem in London, and something that preoccupied John Snow a lot.

So in the mid-19th century, London faced repeated deadly cholera outbreaks.

And germ theory wasn't yet accepted at that time. Miasma theory was still dominant: the theory that diseases were caused by foul air, really bad smells in the air.

And this was even supported by some data at the time. Some people had noticed that there were fewer cholera cases at higher altitudes, where they figured the air must be cleaner.

But Snow wasn't convinced by this.

At that time, drinking water came from the Thames River in London, which was heavily contaminated from sewage. And they were getting their drinking water from the same place they were disposing of things.

So Snow suspected that this was the true source of the cholera outbreaks. And in order to prove that, he began mapping cholera cases with the location of water pumps in London.

So what did he find? Living near pumps that drew their water from the more polluted lower Thames, so downriver, where it was more polluted, greatly increased cholera risk.

And this water served both poor and rich areas across London. It seemed to explain otherwise random outbreaks. The outbreaks of cholera were affecting all kinds of people in all kinds of areas of London.

But water from the more contaminated lower Thames seemed to explain them.

And during a severe outbreak, Snow traced nearly all cases to the Broad Street pump, supplying contaminated water.

Even more convincingly, there was a workhouse and brewery nearby the Broad Street pump that saw almost no cases. And this was explained by the fact that in the workhouse, they had their own well, and in the brewery, they just drank beer.

So he was able to very convincingly show that it was drinking water from the Broad Street pump that explained whether or not you were caught up in the cholera outbreak.

And some cases in surrounding areas far away from the pump turned out to be from people who were traveling all the way to the Broad Street pump because they liked the taste of the contaminated water.

So this is really a demonstration of the kind of thinking that we still use to see if an exposure explains a negative health outcome: linking cases with the exposure, linking the exposure to higher rates than in groups that aren't exposed, and ruling out other explanations. So it wasn't whether you were rich or poor, for example. There was no other explanation left that could explain cases as well as the contaminated water.

So now let's look at two very important and related concepts in epidemiology, which are incidence and prevalence.

So, counting health events is the most fundamental exercise in epidemiology. Before we look at determinants or anything else, we just need to know how many people are diagnosed with a condition, how many deaths are happening.

And then we can look at things like what age groups are suffering most, gender, socioeconomic status, area you live in. But we have to have numbers, accurate numbers, before we can do any of that.

And comparing change over time and between groups is essential to identifying if risk is raised or lowered.

But a raw count doesn't account for population differences. So, ten cases of cholera just in my little village would be very, very concerning, but ten cases in a country of 200 million people wouldn't be as alarming as ten cases out of 1,000. So we always have to have some denominator. We always have to consider the number of cases relative to the population at risk.

So we have lots of measures of disease or death. But all these metrics take basically this format, where the number of cases is the numerator, and the population at risk is the denominator. I don't know if a number is big unless I know how big the population is that that number came from.

So, incidence and prevalence are really the core numbers that we use to keep track of health events.

So, prevalence refers to the current active cases, or the number of active cases during a set period. So, how many people have the condition now as we speak, or over the past week.

Incidence is new-onset cases in a given period. So, how many people are newly diagnosed this week?

So prevalence goes up only when incidence is higher than the combined rate of recovery and mortality. So if you think of the question, “How many people have the condition right now?”, well, I can use the number of new cases to figure that out, but I need to know the duration of the disease as well.

If someone was newly diagnosed two weeks ago, do I think they still have the condition now, or are they still a prevalent case? So I need to consider existing cases passing away, or recovering from the condition, to know prevalence.

So there's a time delay between a change in incidence and a change in prevalence, because today's cases may be tomorrow's recoveries. There's that lag in time between someone being diagnosed and someone passing away or recovering.

So you may be thinking about rates of change and some mathematical ideas, but we use this image here on the right, the epidemiologist's bathtub, to explain this relationship. So in the epidemiologist's bathtub, the amount of water in the bath represents the prevalence, how many cases we have.

And the prevalence lowers because people are recovering, so we represent that as steam rising from the bath. And the prevalence also lowers as people die, so we represent mortality as a leak out of the bottom of the bathtub.

Then incidence is the water coming in from the top, filling up the bathtub.

So if we're leaking prevalence through mortality, and losing prevalence through recovery faster than the incidence tap is refilling the bathtub, then prevalence will go down. If the incidence rate matches the combined recovery and mortality rate, we get a more stable prevalence. And then, of course, prevalence will go up if incidence goes up faster than people recover or pass away.

So both numbers are important to consider: what's the current burden of a disease, the current prevalence, and what's that incidence rate like? Are we expecting prevalence to go up or down in the immediate future?

Now, let's look at exposure and risk.

So, in epidemiology, we try to track if an exposure is associated with increased risk of a negative health outcome. So, a group of people exposed to a chemical: are they at higher risk of a disease?

This allows planning for future healthcare provision and targeting of control measures to mitigate harm. So who's most at risk? Who should we focus on? And who's being exposed now to something that makes them more at risk of a condition in the future? And can we plan for that?

So if we find, for example, exposure to poor air quality is associated with reduced cognitive performance, we might examine causes and attempt interventions to improve this, once we know the association. We might also examine if some groups are more at risk than others: rural or urban, based on class, sex, anything you can think of.

So in Module 1, we looked at relative risk, but I'm bringing it up here again because it's so important in epidemiology. It allows us to consider if an exposed group are more at risk of a negative health outcome than another group, and it reduces to quite a simple number.

So, we typically compare the proportion of people exposed to a hazard who have the negative health outcome with those who are not exposed, and we call this relative risk.

So, if the rate of the condition is the same in the exposed group as those not exposed, they have the same risk, the relative risk is one. If they have lower risk, their relative risk comes out less than one, and if indeed the exposure is associated with increased risk of a health outcome, then the relative risk of the exposed group will be greater than one. So it's a very useful, very interpretable measure.

Because the first thing to ask is: is this exposure really associated with a worse health outcome? Then we know if we need to target and look for causes and come up with health interventions.

Now, the last thing I want to consider in the context of epidemiology is causal inference. So, how do we know something is a determinant? How do we know the exposure was the cause of the greater risk of an outcome?

So it's one thing to say that an exposure raised the risk of an outcome, but it is another thing entirely to say that an exposure caused an outcome.

So if the exposed group and the non-exposed group were identical in every way, except for the exposure, it would be pretty easy to conclude that any differences later on in the health outcomes were caused by the exposure. It's the only thing that explains the difference. It's the Broad Street water, the only difference between people who get cholera and don't is the water they're drinking. There's no differences in how rich they are, that kind of thing. So then it's much easier.

But it's very rare that we have two groups who are truly identical.

So we can either create two identical groups by randomly assigning people to an exposure. Then any differences between the groups are simply random.

But a lot of the time, we're not allowed to control who gets exposed to something, especially if it's a negative exposure.

So a lot of the time, we try to control, as we say, statistically for any differences. So thinking back to regression in Module 1, how we could try and account for a third variable that could influence the relationship between two variables, an exposure and an outcome.

So what are statistical controls? Well, let's take a simple example. Let's imagine my exposed group are lower income than my non-exposed group. So I'm worried that the real reason my exposed group have poorer health outcomes is this lower income. Of course, it's easy to imagine that there might be lots of reasons lower-income people would have poorer health outcomes other than the exposure I'm interested in.

So I could look at bands of equal income. I could look at people in the exposed group and people in the non-exposed group who are in the same income level and see if there's still a difference.

That would be an attempt to control statistically for income. So there's more sophisticated statistical ways to do it, but that's the essential idea: to just compare like with like if overall my groups aren't that similar on something important, like income level.

Now, a useful tool for identifying these third variables that could bias my result—we call them confounds—and a useful way to identify them is with directed acyclic graphs. So we use these in epidemiology to describe causal relationships, and then we can decide what to control for.

So each variable, like an exposure, a treatment, that gets its own node, the circle things. And if we believe there's a causal relation, we draw an arrow from the cause to the effect, to the outcome.

So this directed acyclic graph says that C causes both A and B, and it says A causes B, but it doesn't cause C. And it says that B doesn't cause anything; it's just caused by the others.

Now, because C causes both A and B, it can cause both of them to be higher or lower. It would bias the effect of A on B. If I think A causes B, and I want to measure how much A causes B, if I don't control for C, I could under- or overestimate the effect of A on B.

So variables like C, that cause both the exposure and the outcome, I always need to control for those.

So, controlling for confounds is very important.

But it's quite difficult to always know all possible confounds to control for. How do I know there isn't another variable I haven't thought of, that I haven't measured?

It's quite difficult.

So that's the beauty of a randomized control trial, is when individuals are randomly assigned to the treatment, or the exposure.

We can say that any differences in the outcome are either due to random chance, or they must be caused by the treatment. There's no other explanation.

Because whether or not you ended up in the treatment or control group was totally random. It wasn't determined by your income level or your sex. We assigned you randomly.

And the rate of random positive findings is known, so we can assign it a probability. If there's a difference between the treatment and control group in the outcome, we can say what probability that occurred through random chance and not because of the treatment, which is very useful. So this ability to control for unknown biases—we don't have to have thought of every confound—that makes RCTs the gold standard in causal inference. It controls for things we haven't even thought might be biasing the effect.

So now let's summarize our key takeaways from what we learned today.

So we learned that epidemiology is the study of patterns of exposure and health events in the population.

We learned about incidence and prevalence, which are metrics that track rates of health outcomes and exposures that help us identify trends and to plan interventions.

We learned about exposure and risk. So we identify factors that increase the chance of a negative health outcome to understand potential causes, and to guide prevention.

We looked at causation and controls. So elevated risk alone doesn't prove causation. We control for the potential biases to determine true causal links.

And we learned about randomized control trials, that well-designed trials minimize bias, and they're the gold standard for testing the effect of an intervention.

So you've made it to the end of this series of videos, but there may be future training videos from ISES Europe, for example on legislation. So, do keep an eye on the ISES Europe website.

We thank you very much for your participation and attention, and congratulations on completing all the videos in this series. You can access the other videos on the ISES Europe website with the link here on the slide.

And if you'd like to read more about epidemiology, there's a couple of resources on the screen. The US Centers for Disease Control and Prevention, the CDC, have some quite useful resources as well.

So, thank you very much.