

# Primer: How to Think About Health Statistics

**Lecturer: David Watkins, MD, MPH**

Before we move into the data on non-communicable diseases, it is worth taking a moment to establish a shared foundation for interpreting health statistics. Two disciplines are particularly relevant here, and they are complementary: demography and epidemiology.

Demography is the study of populations. Three elements are relevant to interpreting health statistics and trends: population size (how many people there are), population structure (whether those people are predominantly young, old, or some mix), and migration (how people move between places). These factors matter when interpreting disease burden statistics. For example, if you are an epidemiologist who wants to know how lethal a disease is, you don't just need to know the number of deaths from that disease; you also need to know the total number of people in the population at risk. Dividing the number of deaths by the number of people gives the mortality rate, which makes it easier to compare this disease across populations with different sizes. Conversely, if you are a health policymaker who needs to plan and budget for NCD services, you don't need to know the mortality rate; you need to know the number of deaths.

Demographic data are also important for interpreting mortality rates. The age structure of the population needs to be considered.

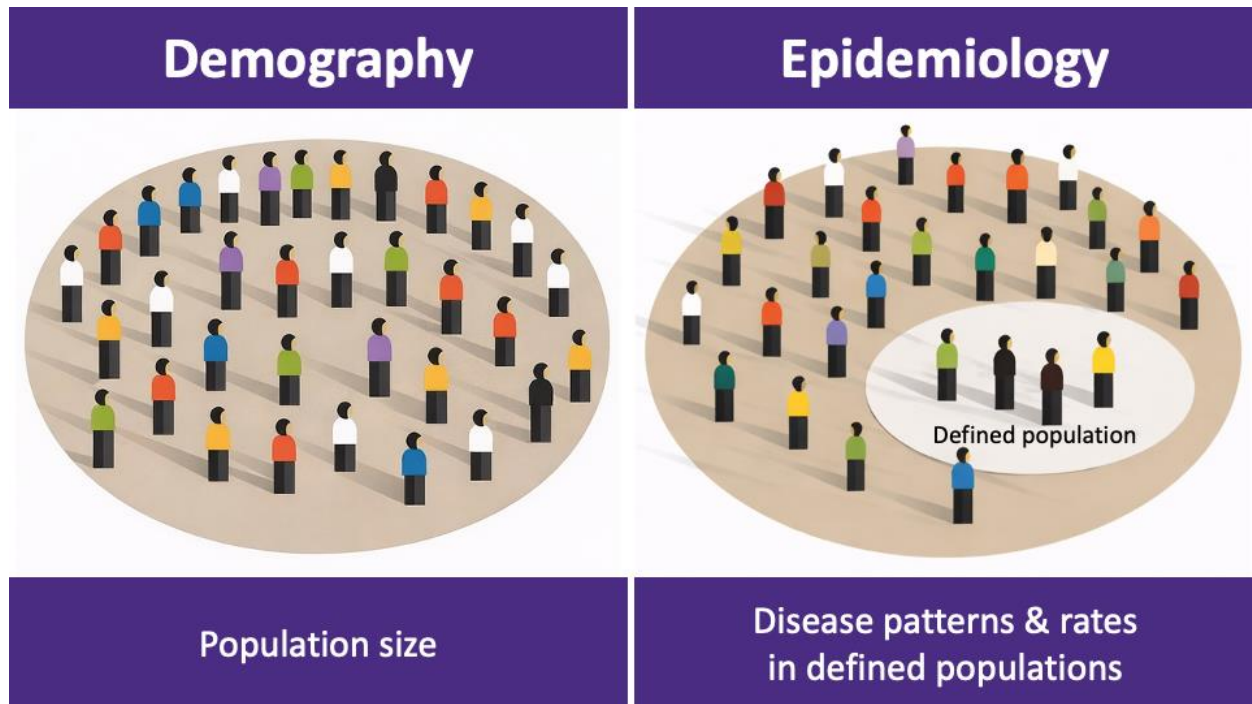
For instance, if a disease kills at a rate of 500 per 100,000 people per year, it matters a great deal whether those people are young or old. Epidemiologists use a statistical technique called age-standardization to compare disease rates in populations with different structures. We will discuss age-standardization later in this lecture.

Migration also affects how we understand the nature of the susceptible population over time, particularly in humanitarian settings where there might be an influx of individuals in poorer health than the general population. For most purposes, though, we focus on how population size and structure shape the statistics we observe around who is dying and from what causes.

Epidemiology is distinct from demography. It is not the study of populations per se, but the study of health-related events and health states. Epidemiologists examine the patterns and frequency of ill health, which we call descriptive epidemiology, as well as the determinants of those patterns, which we call analytical epidemiology. This work is typically conducted in defined populations, using approaches such as case-control studies or cohort studies. Demography, by contrast, tends to focus on the population as a whole.

When we discuss global burden of disease statistics in this course, we are largely working within descriptive epidemiology. However, that work intersects with demographic questions about population size and structure. As you work through the data, it will be useful to keep those two

dimensions separate: how many people are there, on the one hand, and what are the rates of disease in a given age group, on the other. We will return to this distinction throughout the course.



## Core Concepts: Counts, Rates, and Ratios

There are several core concepts in health statistics that will help you interpret data throughout this course. They are conceptually distinct, and it is worth being clear about each one.

A count is simply the number of people with a given characteristic in a population. For example, there were approximately 480 million people living with diabetes in 2017.

A ratio expresses the number of people with a given characteristic relative to some reference group. It is roughly equivalent to a proportion. Using the same example, approximately 6 percent of the global population was living with diabetes in 2017. That is the same information expressed differently. Note that ratios and prevalence figures are sometimes expressed per 1,000 or per 100,000 rather than as a percentage, depending on how common a condition is. It is important to be comfortable reading these interchangeably.

A rate is the number of people with a given characteristic over a defined reference time period, expressed in terms of person-years at risk. For example, the rate of new-onset diabetes in 2017 was approximately 300 per 100,000 person-years. Counts and rates convey related but different information: the rate describes the pace at which events occur, while the count describes the total number of people affected at a given point.

| COUNT                                                                                          | RATIO                                                                                            | RATE                                                                                |
|------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
| Number of people with a given characteristic in a population                                   | The number of people with a given characteristic relative to some reference group                | Number of people with a given characteristic over a defined reference time period   |
|                                                                                                | Roughly equivalent to a proportion                                                               | Expressed in person-years at risk                                                   |
| <b>EXAMPLE:</b><br>“There were approximately 480 million people living with diabetes in 2017.” | <b>EXAMPLE:</b><br>“Approximately 6% of the global population was living with diabetes in 2017.” |  |

Moving fluently between counts, rates, and ratios is an essential skill for interpreting health statistics and seeing the big picture in the data.

## Incidence, Prevalence, and Mortality

Three additional measures appear frequently in this course: incidence, prevalence, and mortality. They are related but distinct.

Incidence refers to new events occurring over a defined time period. It is measured as a risk and is most straightforward to interpret for acute conditions. For example, incidence is a natural measure for heart attacks, where a discrete event occurs at an identifiable point in time. Incidence is typically expressed per 100,000 per year, or per 100,000 person-years.

Prevalence refers to the proportion of a population that has a given condition at a particular point in time. It is expressed as a ratio, often as a percentage or per 100,000. Prevalence is particularly useful for describing chronic, stable conditions, which is why it is used so often in the context of non-communicable diseases.

Mortality refers to deaths occurring over a defined time period. Like incidence, it is expressed as a rate or risk, typically per 100,000 population, or per 100,000 person-years.

| INCIDENCE                                                | PREVALENCE                                                                        | MORTALITY                                                                                                         |
|----------------------------------------------------------|-----------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------|
| New events occurring over a defined period of time       | Proportion of population that has a given condition at a particular point in time | Deaths occurring over a defined time period                                                                       |
| Measured as a risk                                       | Expressed as a ratio, often as a % or per 100,000                                 | Expressed as a rate or risk, typically per 100,000 population, or per 100,000 person-years for very common events |
| Used to interpret acute conditions (e.g., heart attacks) | Used to describe chronic, stable conditions, often in the context of NCDs         |                                                                                                                   |
| Expressed as per 100,000/year or in person-years         |                                                                                   |                                                                                                                   |

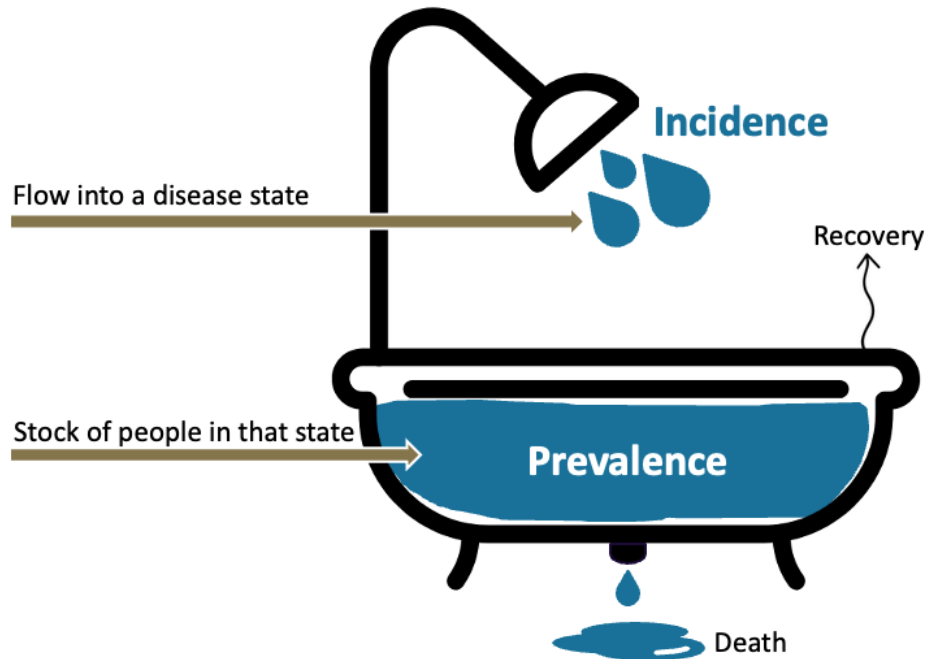
## Why Prevalence Is Typically Higher Than Incidence for Chronic Conditions

Understanding these definitions raises an important question: how do these measures relate to one another in practice?

For most chronic non-communicable diseases, prevalence is substantially higher than incidence. This is an important and recurring pattern in the data.

The reason is straightforward. Incidence captures only the new cases arising in a given year. Prevalence captures everyone currently living with a condition, including people who developed it in prior years and have continued to live with it. Because people with chronic conditions do not typically recover or die immediately, the pool of prevalent cases accumulates over time. As a result, prevalence for many NCDs is often an order of magnitude higher than incidence.

To use an analogy from systems engineering: incidence is the flow into a disease state, while prevalence is the stock of people in that state. When people enter slowly but stay for a long time, the stock grows large relative to the flow.



Adapted from: Bunce & Wormald, 2024

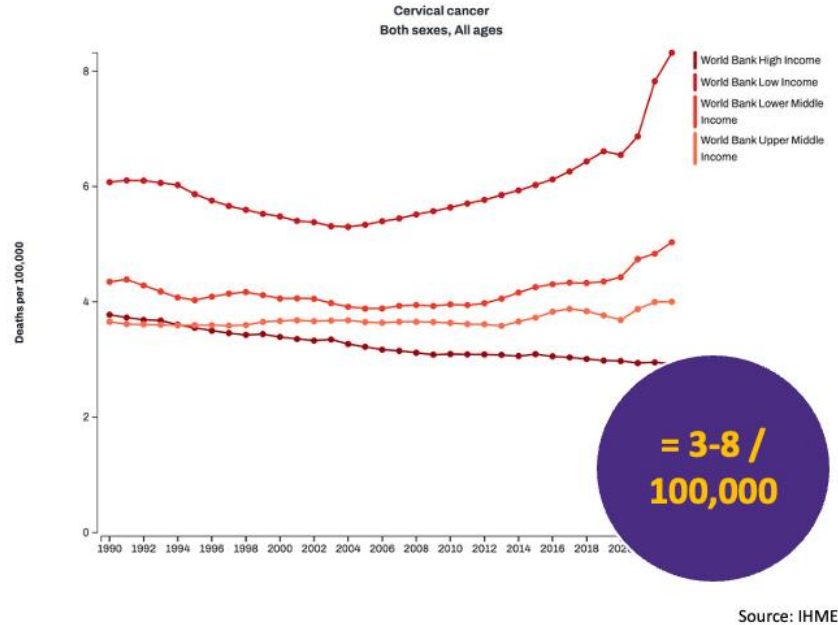
## Why Mortality Is Lower Than Prevalence

A related pattern emerges when we compare mortality to these measures, and it is equally important for interpreting NCD data.

Death rates for most chronic conditions are lower than prevalence estimates, often substantially so. This is because not everyone living with a chronic condition dies from it in any given year. For conditions that are not uniformly or rapidly fatal, you would expect the death rate per 100,000 population to be considerably lower than the prevalence rate.

To illustrate this with approximate figures from cervical cancer data: incidence might be in the range of 8 to 20 per 100,000, prevalence in the range of 40 to 100 per 100,000, and death rates around per 100,000.

### Cervical cancer death rates



These differences reflect the fact that people live with the condition for some time after diagnosis before dying. Keeping this ordering in mind will help you read statistics accurately and avoid common misinterpretations.

### Age Patterns and Age Standardization

So far, we have focused on how measures are defined, but we also need to how population structure shapes what those measures show.

The age distribution of a population has a significant effect on the crude mortality statistics we observe, and this creates a common source of confusion when comparing countries.

Consider two populations with different numbers of individuals and different age structures.

## Age patterns of outcomes

|              | Country A     |                  |                          | Country B     |                  |                          |
|--------------|---------------|------------------|--------------------------|---------------|------------------|--------------------------|
| Age group    | Deaths        | Population       | Death Rate (per 100,000) | Deaths        | Population       | Death Rate (per 100,000) |
| 0 to 14      | 300           | 300,000          | 100                      | 60            | 100,000          | 0                        |
| 15 to 49     | 900           | 400,000          | 225                      | 500           | 600,000          | 100                      |
| 50 to 69     | 9000          | 200,000          | 4500                     | 8000          | 900,000          | 2000                     |
| 70 or older  | 6000          | 100,000          | 6000                     | 11,000        | 400,000          | 4333                     |
| <b>Total</b> | <b>16,200</b> | <b>1,000,000</b> | <b>1620</b>              | <b>19,560</b> | <b>2,000,000</b> | <b>2120</b>              |

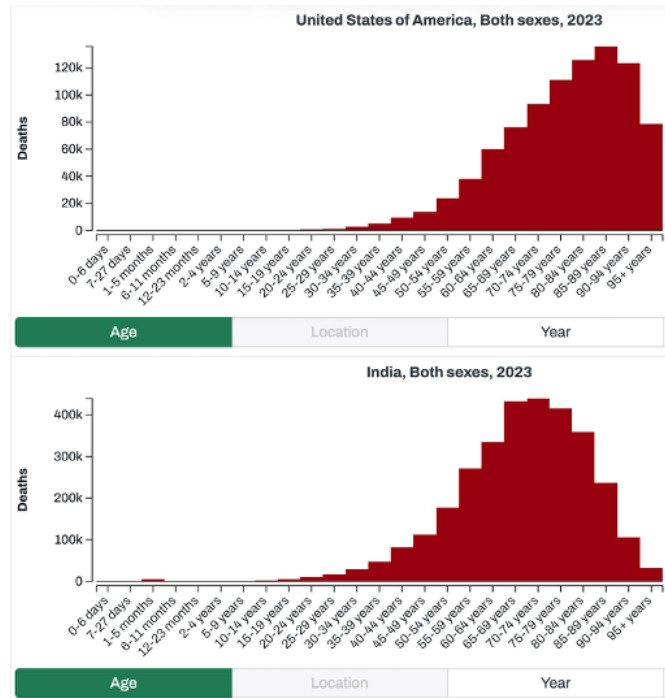
Country A has a younger population structure, with a large proportion of children and working-age adults. Country B has an older population structure, with a larger share of people aged 65 and over. If you compare the crude death rates of these two countries, you may find that Country B has a higher crude rate, simply because older people die at higher rates regardless of where they live or what health system they have access to. You might incorrectly conclude that Country B is performing worse on health outcomes.

When you examine age-specific mortality rates, however, the picture often reverses. Country A may actually have higher death rates at younger ages, including children and working-age adults, suggesting that its population is experiencing preventable mortality that Country B has reduced. The difference in crude rates is an artifact of the age structure, not a reflection of true differences in health system performance or disease burden.

This is why age-standardized mortality rates are used when comparing countries or tracking trends over time. Age standardization adjusts the observed rates as if both populations had the same age structure, allowing for a fair comparison. This is often described as comparing apples to apples.

A useful real-world example of this dynamic is the comparison between the United States and India. The crude death rate in the United States is higher than in India, which might seem counterintuitive.

## Age distribution of cardiovascular disease deaths: United States & India, 2023



Source: IHME

However, when you look at age-specific rates, particularly for cardiovascular disease in middle age, India shows considerably higher death rates. The United States has an older population, so its crude rate is elevated by deaths among the elderly. India has a younger population overall, but is losing people at earlier ages, which is the more meaningful indicator of preventable mortality. Age standardization reveals this distinction.

This pattern appears repeatedly when examining NCD data across countries and over time. The age dynamics of non-communicable diseases are a critical part of the story, and you will encounter this concept many times throughout the course.

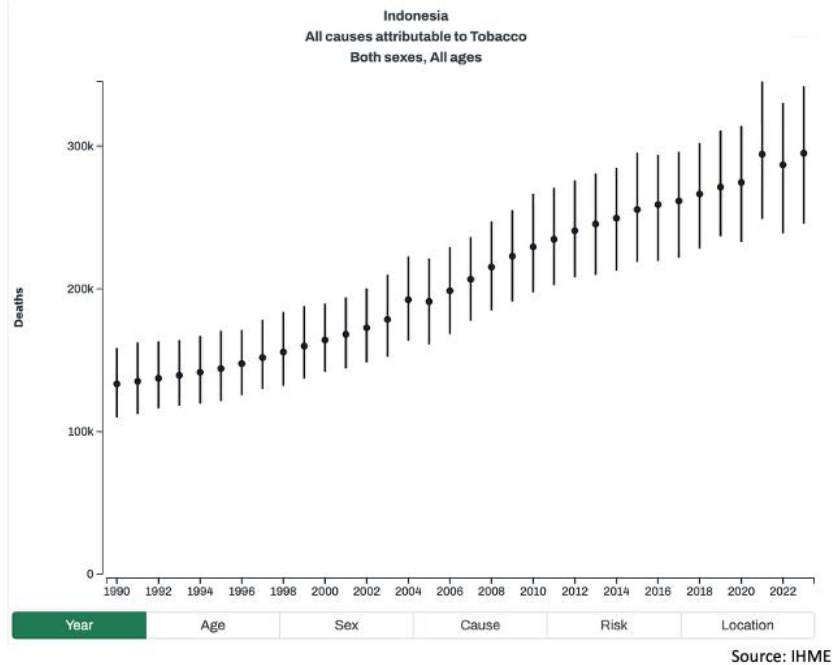
## Risk-Attributable Mortality and Population Attributable Fractions

A further concept that is central to interpreting NCD statistics is risk-attributable mortality. This refers to deaths from a specific cause that can reasonably be attributed to a particular risk factor.

The underlying approach draws on analytical epidemiology. From studies such as cohort studies or case-control studies, we can estimate that a given risk factor, such as tobacco use, increases the risk of dying from a condition such as cardiovascular disease by a certain percentage.

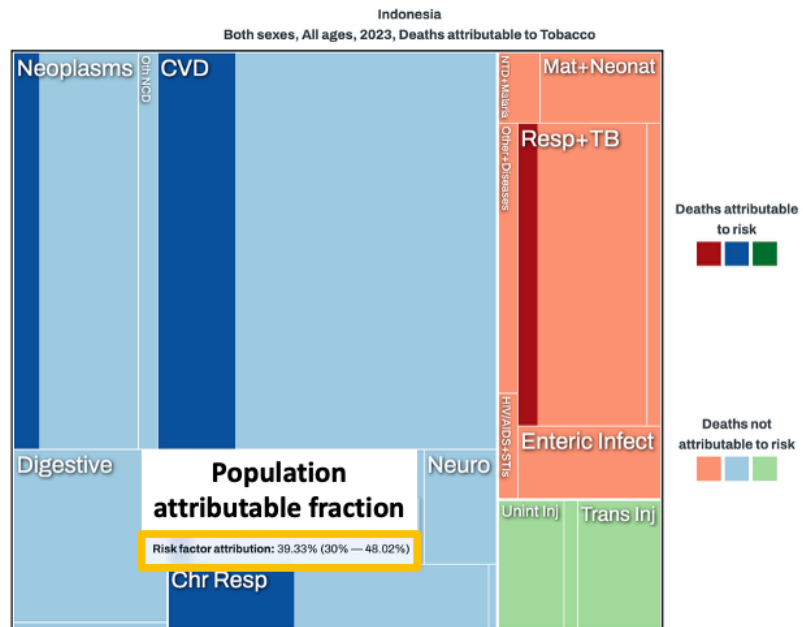
## Noncommunicable Diseases in Global Health

Number of deaths attributable to tobacco, Indonesia, 1990–2023



If we also know how prevalent that risk factor is in a population, and how many deaths from that condition occur, we can calculate the proportion of those deaths that are attributable to that risk factor. This is known as the population attributable fraction.

Causes of death attributable to tobacco: Indonesia, 2023



An extension of this approach allows us to compare observed mortality to what would theoretically occur if the risk factor were at an ideal, or counterfactual, level. The difference

between observed deaths and deaths at that ideal level represents the excess mortality attributable to the risk factor. This is the methodology used, for example, to estimate the number of deaths worldwide attributable to tobacco use, obesity, dietary factors, or high blood pressure.

It is important to understand that these figures are estimates based on triangulation of data sets. We do not directly observe and record tobacco-attributable deaths in routine statistics. Instead, we combine data on how many people died of various causes, how common tobacco use is in that population, and what the excess risk of death is for tobacco users, and we calculate an estimate from those inputs.

### **Important Limitations of Risk-Attributable Estimates**

There are two limitations of this approach worth keeping in mind as you interpret these statistics.

First, the choice of reference level matters. For tobacco, the reference is typically no tobacco use, which is straightforward. For other risk factors, such as salt consumption, dietary patterns, or body weight, the ideal level is more contested and can significantly affect the resulting estimates. When reading risk-attributable mortality figures, it is worth asking what reference level was used and whether that assumption is defensible.

Second, when you add up the deaths attributable to multiple risk factors, the total will exceed 100 percent of observed deaths. This is not an error. It reflects the fact that risk factors overlap and interact. A person who dies of cardiovascular disease may have been a smoker, had a poor diet, and been physically inactive. Each of those risk factors contributes to the attribution calculation independently. The method is a post-hoc counterfactual analysis, not a direct causal accounting of each death. Keeping this in mind will prevent confusion when you encounter figures that appear to sum to more than the total number of deaths.

### **The Global Burden of Disease Study**

Much of the data referenced in this course comes from the Global Burden of Disease Study, or GBD. It is worth understanding what this study is and how it is produced.

The GBD Study began as a World Bank and World Health Organization initiative in the 1990s to systematically quantify causes of death and disease globally. Over time, that enterprise has grown considerably and is now housed at the Institute for Health Metrics and Evaluation at the University of Washington. The World Health Organization produces its own parallel statistics, known as the Global Health Estimates, which use broadly similar methods and data sources. Both are credible resources, and either can be used appropriately, though there are methodological differences between them.

The GBD Study is a systematic attempt to estimate the causes of death, disease, and injury across all countries and territories, as well as the contribution of major risk factors.

Current estimates cover 204 countries and territories,

- approximately 292 causes of death,
- 375 causes of disease,
- and around 88 risk factors.

These estimates are produced by synthesizing a wide range of data sources, including death records, verbal autopsy data, police reports, risk factor prevalence surveys, and other inputs. It is a large and complex data enterprise.

### **The GBD Cause Hierarchy**

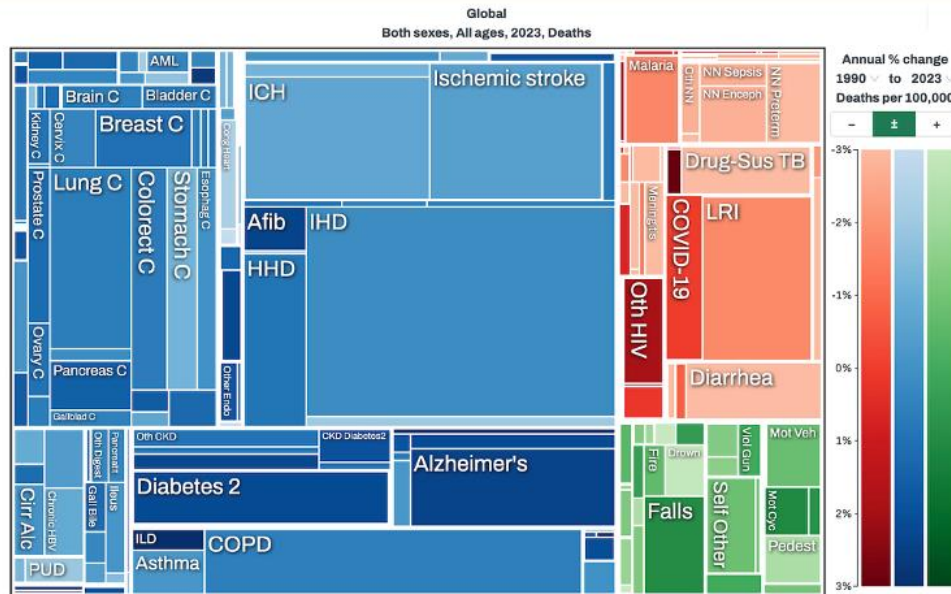
To interpret these data correctly, we also need to understand how causes of disease and death are organized within this system.

The GBD Study organizes causes of death and disease in a hierarchy, moving from broad categories down to specific conditions. At the highest level, all causes of death are divided into three groups.

Group 1 covers communicable, maternal, neonatal, and nutritional diseases. Group 2 covers non-communicable diseases. Group 3 covers injuries. Within each group, causes are progressively disaggregated. For example, within non-communicable diseases, one category is cancers. Within cancers, there are approximately 49 specific cancer types, each defined by anatomical site, such as breast cancer, lung cancer, and colorectal cancer. Similarly, cardiovascular diseases is a category that contains multiple distinct conditions, each with its own epidemiology, but grouped together for convenience and comparability.

This hierarchy is the structure behind the visualizations you will encounter when working with GBD data, including the tree maps that show causes of death disaggregated from the broadest level down to specific conditions. Understanding the hierarchy will help you navigate those tools and interpret the data accurately.

This hierarchy is the structure behind GBD data visualizations



## Summary

In summary, interpreting health statistics well requires familiarity with a small set of foundational concepts. Demography and epidemiology are complementary disciplines: demography helps us understand population size and structure, while epidemiology focuses on patterns of disease and their determinants. Counts, rates, and ratios each convey different information, and moving fluently between them is essential. Incidence measures new cases over time, prevalence measures the existing burden of disease in a population, and mortality measures deaths over a defined period. For chronic conditions, prevalence is typically much higher than incidence, and death rates are lower than both, because people live with these conditions for extended periods.

Age structure has a profound effect on the crude statistics we observe, and age standardization is necessary for meaningful comparisons across countries or over time. Risk-attributable mortality estimates allow us to link deaths to specific risk factors using population attributable fractions, but these figures are model-based estimates with important limitations, including the choice of reference level and the fact that attributable deaths across multiple risk factors will sum to more than the total. Finally, the Global Burden of Disease Study is the primary data source for the statistics discussed in this course, providing systematic estimates of causes of death, disease, and risk factors for 204 countries and territories.