<1>Chapter 1: Fundamentals

In this Chapter, we define and provide examples of several key terms used in public health and healthcare modeling research. We begin by clarifying the differences between key terms used to describe rates of disease (incidence, prevalence, and mortality) as well as the performance characteristics of tests used to detect disease (sensitivity, specificity, positive predictive value, and negative predictive value), prevent or treat disease (odds ratios, relative risks), understand studies (case-control, cohort, and randomized controlled trials), and avoid common study problems (bias, confounding).

<2> Incidence, Prevalence and Mortality: An Example From HIV in Uganda

In the 1990s, HIV (human immunodeficiency virus) ravaged entire countries, and debate ensued regarding the most appropriate strategy to reduce transmission of the disease. Public health experts advocated for wider distribution of condoms to prevent sexual transmission of HIV, but officials at the United States White House (under then-President George W. Bush) argued that such an approach would legitimize unsafe sex. Instead, White House officials declared that "abstinence-only" educational programs were effective for preventing HIV infections. They reported that, when broadly instituted, abstinence-only programs were shown to generate substantial reductions in the number of people with HIV over time.

Public health experts were skeptical, but they did not find any problems with the data collection that led to the reports released by the White House. The number of HIV-positive individuals had been studied appropriately by surveilling a similar population of people over many years, in areas without major migration, and in locations where other factors contributing to HIV transmission (the local economy, for example) had not changed over the study period. If the data were in fact correct, what reason—other than success of abstinence-only programs—might explain why the number of people infected with HIV had declined after the introduction of the programs in Uganda?

To answer this question, we need to clarify three key terms used to describe rates of disease: incidence, prevalence, and mortality. Perhaps the most straightforward way to remember the important distinction between these terms is to visualize a bathtub, as shown in Figure 1.1.

[INSERT FIGURE 1.1 HERE].

If the bathtub reflects the state of HIV in an area, then we can conceptualize the *flow* of water into the bathtub as the *incidence* rate of HIV, or the number of people over a period of time (per week, per month, or per year, to give a few examples) who are newly infected with HIV in the studied Ugandan communities.

We can conceptualize the *level* of water in the bathtub at the *prevalence* of disease, or the number of people at a particular time of a study who are infected with HIV. Unlike a rate of incidence, the prevalence of a disease reflects a single point in time—not a change in people over a period of time. The statistics used by the White House were prevalence statistics. In other words, White house officials compared the number of people with HIV in the years prior to the abstinence-only education program to the number of people with HIV in the years after the abstinence-only education program.

Finally, we can conceptualize the *flow* of water out of the bathtub as the *mortality* rate from HIV, or the number of people over a period of time who have died from HIV or HIV-related complications.

As illustrated by the bathtub analogy, incidence and mortality are *rates* of disease—they indicate the number of newly-infected or dead persons per unit time. In contrast, prevalence is a *level* of disease (the number of people with the disease at the moment of measurement). Using these definitions, can you determine how White House officials may have incorrectly ascertained the effectiveness of abstinence-only programming?

There are two reasons why the level of water in a bathtub might decrease. First, the faucet could be turned down, causing less water to enter the bathtub through the faucet than leaves the bathtub through the drain, per unit time. Alternatively, water could exit the bathtub more rapidly (suppose a plug is removed from the drain, for example) such that more water leaves the bathtub through the drain than enters the bathtub through the faucet, per unit time.

Similarly, the prevalence of a disease might decline for one of two reasons: either less people are getting the disease than are dying from the disease (the incidence rate has declined relative to the mortality rate), or more people are dying from the disease than are getting the disease (the mortality rate has increased relative to the incidence rate).

Unfortunately, in the case of HIV in Uganda, the incidence rate of HIV had not declined as a consequence of effective abstinence-only programs. Instead, scarce opportunities for treatment caused the mortality rate to continue increasing in the population. Hence, the diminishing prevalence of HIV-positive individuals in Uganda during this time did not reflect the effectiveness of abstinence-only programming, but the rapid and deadly decline faced by patients that contracted HIV [1].

<2> Sensitivity, specificity, and predictive values: An example from lupus screening

The model of allocation processes that we derived from our example of *Salmonella* contamination was based on a key assumption: that when we allocate resources such as inspectors to a factory, our resources will be perfectly effective: every inspector will find every contamination event at their assigned factory.

Reality, of course, is far less than perfect.

When evaluating screening strategies, we must often decide between imperfect tools. For most medical and public health situations, we have to carefully quantify how far our tools are from perfection, to ensure that we use our tests and tools as well as possible.

Suppose we take the example of screening patients for Systematic Lupus Erythematosus (SLE), a rare disease affecting about 0.2% of the population, and commonly known as “lupus.” Lupus is an autoimmune disease, meaning it is triggered by self-reactive antibodies that attack a person’s own organs, joints, and skin. Unfortunately, many people who have lupus are not diagnosed until their antibodies have already caused severe and irreversible damage. If caught in the earlier stages of disease, however, people with lupus can receive medical treatment that can prevent serious complications.

The Lupus Foundation of America started a campaign to encourage people with symptoms characteristic of lupus to contact their medical providers and request screening. Billboards, radio and television advertisements, as well as social media campaigns, were started to encourage young women to receive a screening test for lupus, known as the antinuclear antibody (ANA) test.

The Foundation argued that the ANA test would be an excellent recommendation for most people, since it was known to have a high capture probability for detecting lupus. However, medical groups worried that recommending the treatment widely would cause a serious problem. To visualize the problem, we can a scenario in which the ANA test is applied to a population of 10,000 people. Typically, such data are portrayed using a *contingency table*, which summarizes not only the number of people who test positive or negative using the ANA test, but also the number of people who are later confirmed to truly have (or not have) lupus.

[INSERT TABLE 1.1 HERE].

From this table, we can observe that the capture probability of detecting a patient with lupus, if they truly have lupus, is 100% [2]. Out of 20 people who had the disease, all 20 test positive. The capture probability *of detecting a person with the disease* is commonly referred to as a test’s *sensitivity*. We can calculate the sensitivity as the number of people testing positive who actually have the disease, divided by the total number of people having the disease.

[EQUATION 1.8] Sensitivity = (true positives)/(all persons with the disease)

A problem with the ANA test, however, is that only 86% of people who did not have lupus ended up having a negative test (8,593/9,980). The probability of *correctly identifying a person who does not have the disease* is known as a test’s *specificity*.

person who does not have the disease is known as a test's specificity.

[EQUATION 1.9] Specificity = (true negatives)/(all persons without the disease)

The low specificity of the ANA test generated many "false positive" results. In other words, 1,397 people who did not have lupus were incorrectly labeled as having lupus based on the ANA test.

We can fill out a more complete contingency table to help us determine the value of a particular medical test; in Table 1.2, we add key terms that are commonly used to label people who appear in each cell of the table and summarize the value of a test in terms of the *positive predictive value* (the probability of having the disease is a person tests positive) and *negative predictive value* (the probability of not having the disease if the person tests negative):

[INSERT TABLE 1.2 HERE].

As shown in Table 1.2, the positive predictive value for the ANA test is very low, so having a positive test does not necessarily mean that a person has lupus. (In fact, patients with positive test results most likely do not have lupus). Conversely, the negative predictive value for the ANA test is 100%, so if a person is negative, they can rest assured that they do not have the disease. Hence, the ANA test may still be useful, but only for ruling-out the disease among people truly suspected of having the illness. For the general population of patients having aches and pains without any more specific symptoms or family history of lupus, the ANA test is more likely to produce a "false positive" result and cause needless worry.

A common pair of acronyms used to remember when tests are useful is SPIN and SNOUT: a SPecific test helps us to rule-IN the disease (diagnose someone), whereas a test with high SeNsitivity helps us to rule-OUT the disease (if a person is negative, they can be reassured that they are unlikely to have it). Tests are rarely highly sensitive and highly specific, so we usually apply them in sequence: first, we use a test with high sensitivity to rule-out many people who don't have the disease, then administer a test with high specificity among the persons who tested positive to diagnose people with the disease.

<2> Risks and study designs

Detecting disease accurately is critical to nearly every public health or health care task, but the natural next step is to provide an effective intervention that reduces the morbidity and mortality caused by that disease. For example, many programs in the United States seek to detect people with "pre-diabetes" (a state of not yet having irreversible diabetes, but being at high risk for eventually developing the disease). Persons labeled as having pre-diabetes may benefit from an effective nutrition and physical activity program that delays or prevents the onset of full-blown diabetes. Similarly, people who have an early stage of disease may be able to receive effective pharmaceutical treatments to prevent complications or death from the disease. How can we know if an intervention is actually effective?

A similar problem arises if we wish to detect whether an environmental agent may be increasing the risk of disease—for example, whether exposure to a chemical might cause cancer, or whether people with certain health-related behaviors (e.g. tobacco smoking) have a higher risk of lung disease.

Just as we created a contingency table to determine how many people with a disease had a positive or negative test result, we can create analogous contingency tables to determine if our interventions improve the chances of living a healthy life or whether environmental or personal characteristics worsen the chances of living a healthy life. In Table 1.3, we illustrate such a contingency table:

[Insert TABLE 1.3 here]

In Table 1.3, the rows refer to whether or not a person was exposed to a chemical, and the columns tell us whether or not these people later got cancer or remained cancer-free. Typically, these data are obtained from a *cohort study* (also commonly called a longitudinal cohort study), in which we follow a group of people over time and evaluate how many of the people exposed to a chemical developed cancer and how many people not exposed to the chemical got cancer.

As shown in Table 1.3, the absolute number of people who didn't get cancer is, thankfully, higher than those who did. Also, as expected, fewer people were exposed to the

thankfully, higher than those who did. Also, as expected, fewer people were exposed to the chemical than were unexposed. How do we quantify the extent to which the chemical exposure is associated with cancer?

Epidemiologists typically quantify the risk of disease given an exposure as a *relative risk (RR)*, sometimes also referred to as a "risk ratio", which is calculated as:

[EQUATION 1.10]

In the case of the example shown in Table 1.3, the relative risk would be [15/(15+70)]/[5/(5+150)] = 5.5, which can be interpreted as suggesting that there is more than a five-fold higher risk of getting cancer if a person was exposed to the chemical than if they were not exposed.

In the case of some rare diseases, it is hard to perform a cohort study because even people who are exposed very rarely get the disease, and only extremely large sample sizes (read: very expensive studies) would be able to capture enough people who eventually develop the rare disease to calculate a relative risk. In this circumstance, epidemiologists typically perform a *case-control study* (also commonly called a retrospective case-control study), which involves finding people who already have the disease and asking them whether they were previously exposed to the chemical or factor under study. These *cases* are compared to people who don't have the disease (*controls)*. The problem with case-control studies is that they only sample groups of people with and without the disease instead of the entire population. Thus, we are unsure how many people in the overall population were actually exposed to the chemical and how many were not. This complicates our ability to determine an accurate denominator in our risk estimates. In case-control studies, epidemiologists will calculate an *odds ratio (OR)* instead of calculating a relative risk:

[EQUATION 1.11]

If Table 1.3 were data from a case-control study, then the OR would be (15/70)/(5/150) = 6.4. Here, we interpret the odds ratio by saying that the "odds" of getting cancer when exposed to the chemical are over six-fold the "odds" of getting cancer if a person is not exposed to the chemical. The term "odds" is deceptive, however, as it is a ratio of two factors in the contingency table, not a probability. For very rare diseases, the odds ratio and the relative risk will be similar values, because the number of people who are exposed and didn't get the disease, and the number of people not exposed who didn't get the disease, are relatively large; hence, these values are similar to the total number of people exposed and the total number who did not get the disease, making Equation 1.11 very similar to Equation 1.10.

If we wish to determine if an exposure results in disease, then cohort and case-control studies are reasonable investigative strategies, especially because it is usually unethical or impractical to subject people to an exposure to determine if they eventually get the disease under study. Conversely, if we wish to determine if an intervention (such as vaccine or pharmaceutical) can help prevent or treat a disease, it is often ethical and practical to subject people to the intervention to determine if it truly mitigates disease risk or disease mortality. The beneficial effects of interventions are often best tested using a *randomized controlled trial* (RCT). An RCT involves randomly flipping a coin and assigning some people to get the intervention (e.g., a new pharmaceutical medication) and assigning other people to a "placebo" group (e.g., an empty pill capsule). The benefit of an RCT is that, in the process of randomly assigning people to the intervention group or comparison group, *confounding factors*—that is, factors correlated with both the intervention and the outcome of interest—are distributed equally into both groups. If confounders are distributed equally between groups, they are washed out from the statistical analysis of the intervention, allowing us to correctly estimate the impact of the intervention on the disease risk or mortality.

For example, suppose that we want to know if a particular vitamin was associated with a lower risk of heart disease. If we just did a cohort study of how many people were exposed to the vitamin and how many got heart disease, we might find data such as that shown in Table 1.4:

[INSERT TABLE 1.4 HERE].

Based on the data from Table 1.4, we might conclude that the vitamin is very effective at preventing heart disease, because the relative risk is RR = [3/(3+80)]/[7/(7+60)] = 0.3, meaning that people who took the vitamin only got heart disease about one-third as much as people not taking the vitamin. However, a confounding factor may account for these results. Suppose that

the vitamin in question is found disproportionately in green vegetables. Perhaps the vitamin leads to lower incidence of heart disease, or perhaps something else in green vegetables is responsible for the lower incidence of heart disease (fiber, for example). Other factors like fiber may "travel alongside" the vitamin and generate the reduction in heart disease, but we have wrongly attributed the benefit to the vitamin instead of to fiber.

An RCT would be able to separate out the impact of the vitamin from the impact of "fellow travelers" (confounders) to determine whether the vitamin is actually producing the benefits observed. An RCT could, for example, be designed so that some people were randomly given a pill containing only the vitamin and other people were randomly given a pill that was just an empty capsule (a placebo). Because we randomly assigned people to the vitamin or placebo pill, the number of people in each group who eat green vegetables should be roughly equal if we have a large-enough sample size—consequently washing out the effect of green vegetables because both groups would have people with high and low green vegetable consumption, and the only effect left would be from the vitamin itself. We could then recalculate the relative risk from the RCT data knowing that the outcome should be just the effect of the intervention (the vitamin) rather than confounding factors (like fiber).

RCTs are not always possible, as in cases when it would be unethical to randomize people to harmful exposures. Unfortunately, medical history does have examples of highly unethical studies, including the infamous Tuskegee Syphilis Study in which Black men in the southern United States were intentionally prevented from receiving treatment for syphilis to understand how the disease affected their bodies. Hence, it is important for us to continue doing case-control and cohort studies to identify how natural occurrences might affect disease risk.

Relative risks and odds ratios will be frequently incorporated into our models, particularly when we seek to simulate the benefit incurred from an intervention or assess how rates of disease may be affected if exposure to environmental or behavioral factors were to change.

<2> Bias

When we gather data to put into our models, we usually have to think about whether those data are *biased*, or producing a false estimate of reality.

Two forms of bias are particularly important for modelers of public health and healthcare systems to be aware of.

The first—*selection bias*—refers to the problem of "comparing apples to oranges," or having one group of people compared to a group of people who are very different from them. For example, a famous study reported that workers in shipyards who had been exposed to radiation many years ago were actually healthier and had lower rates of death than people who did not get exposed to radiation [3]. At face value, the study seems likely to be wrong—but why? Further investigation revealed that the people who conducted the study had compared workers in the shipyard to a broad population of people. Importantly, workers in the shipyard were only selected from a relatively healthy population of young men who had passed an initial physical exam. By contrast, the shipyard workers were compared to a broad population including very elderly people. Hence the study was biased, and a proper comparison would have been between young healthy men who entered the shipyards and young healthy men who did not.

Another example of selection bias was observed in recent studies of nutrition. A study compared obesity rates among people who received "food stamps," a form of financial assistance for low-income populations in need of food. A study compared the food stamp recipients to non-recipients, and found that persons using food stamps had higher rates of obesity. The authors concluded that food stamps themselves led to obesity. However, the people receiving food stamps tended to live in neighborhoods that had higher concentrations of fast food restaurants and fewer grocers selling fresh fruits and vegetables. Hence, selection bias could have occurred, and a fairer comparison would have been between food stamp recipients and non-recipients in similar neighborhoods [4,5].

A second form of bias is known as *information or observation bias*, sometimes also called misclassification bias. Information or observational bias refers to a problem that arises when an exposure or outcome is measured incorrectly. This type of measurement error can affect both groups equally. For example, suppose a laboratory test is used to assess a person's

cholesterol levels to see if cholesterol relates to heart disease. If the laboratory test underestimates everyone's cholesterol values by the same amount, it can produce *nondifferential misclassification*—meaning that people with both high and low cholesterol values are equally affected. This is nevertheless a problem, because we will not know the true association between correctly-measured cholesterol levels and heart disease.

Differential misclassification—when measurement error is greater among one group than another—poses an even greater problem. A classical example of differential misclassification is when people who are affected by some exposure tend to remember having a problem more than people who were not exposed. In a study of the chemical Agent Orange (a substance used by the U.S. military in the Vietnam War), for example, pilots in the U.S. Air Force who had been exposed to the chemical remembered having skin rashes after exposure much more than pilots who hadn't been exposed; but objective physical exam records show that the two groups of pilots actually had the same rate of skin rashes. Pilots exposed to Agent Orange were more attuned to their rashes (they remembered them more readily) than those who were not exposed [6]. Another example of information bias occurs with mothers who have children with congenital birth defects. Mothers whose children are born with birth defects recall being exposed to medications more than mothers whose children are born without defects, even if they took the same number of medications [7].