CHAPTER 1

Structures for Asking Questions Epidemiologic Methods Can Answer

LEARNING OBJECTIVES

By the end of this chapter, you will be able to:

- Discuss the importance of methods in research
- Develop research questions
- Construct null and alternative hypotheses
- Discuss the importance of operationalization of variables
- Develop a research question matrix
- Define the counterfactual and referent and the roles they play in epidemiologic research
- Discuss the concept of causality

Why Methods Matter

People have a natural tendency to look first to the "facts." In public health, those facts are often statistics: How many individuals have tuberculosis in the region? How many people have high blood pressure as defined by a particular cutoff value? How many individuals acquire HIV annually? Answers to these questions provide immediate, necessary, and useful information. For many, this may be enough. For others, this may be only the beginning. How this information is ascertained relates directly to what the facts are. Facts, numbers, and statistics are direct extensions of the methods that were used to collect them. To understand the underlying method is to go deeper, to understand how the statistics are right and how they are limited by the way they were collected. For those interested in fields that use epidemiology—and that includes virtually every public health discipline—it is essential to understand the underlying methods of this foundational science of public health. This understanding enhances not only one's ability to use the statistics while implementing or analyzing studies but to design studies capable of yielding valid information.

It is easy to consider methods as an abstract rather than as an applied concept. Yet nearly every activity one performs, from the most pedestrian to the most complex, has a method implicit in it, demanding a systematic approach to conduct. Consider cooking, an activity one does routinely and that demands a careful execution of steps. In a recipe, the methods are documented and clearly laid out. In most instances, informed or minor deviations to the recipe result in a product similar to one that would be produced following the original recipe. However, under certain conditions, even the most minor of deviations can result in disaster. Imagine, for example, a cake without the requisite leavening or eggs. Many of us are also familiar with the demanding methods entailed in fields such as physics or chemistry, which require a lab notebook, an explicit set of instructions with which to conduct an experiment, and documentation of each and every step performed. To conduct an experiment, exactitude in documentation of the steps is essential. How much of the buffer was titrated? What was the specific and quantified result of the experiment? No matter what the context, precision in the system of action and the documentation is what makes the experiment possible, interpretable, and repeatable. Similarly, our research protocols allow us to implement studies in a valid and replicable way over time.

The level of detail in our methods and requirements for documentation are similar to those used in other sciences such as biology or chemistry. We use our methodological tool kit to answer questions. We document these methods so that they are replicable and testable by others and also so we can assess and measure deviations in the methods themselves as well as in the subject under study. How data are collected has a profound effect on the ultimate results. *Methods matter.* Looking at results in the absence of the context in which they were derived limits our ability to appropriately interpret the results. To understand the data we must understand how those data were collected.

The purpose of this chapter is to start you on your path to viewing information through a methodological lens: considering data, information, statistics, and qualitative contexts and the critical role of the methods used to obtain them. Once this foundation has been established, it will be easier to understand the emphasis placed on methods. Having this foundation will allow you to move beyond the facts about a specific health condition and gain an understanding of what the condition is, what independent variables (potential risk factors) are associated with it, what confounders (characteristics that are non-causally associated with the condition) may exist, and how a given outcome can be prevented or treated. Whether conditions are new, emerging, or reemerging, chronic or acute, infectious or noninfectious, understanding which methods should be employed in a study—and those methods' relative strengths and limitations—makes all the difference in informing our understanding of determinants of disease and development of public health interventions to improve the overall health of populations. In this chapter, we will look at several examples that describe how to dig deeply into methods, thus laying a foundation for developing your conceptual understanding of the importance of why methods matter.

An Example of the Importance of Methods in Epidemiology

We are exposed to some amount of media regarding medications nearly every day. This might be at our own healthcare provider's office or through reading a magazine, seeing a commercial, or using an over-the-counter drug. Most newer medications have undergone a rigorous examination through clinical trials—studies that systematically review the safety and efficacy of the medication in humans. One level of methods is readily apparent here: the epidemiologic methodology required to develop, implement, analyze, and interpret data from the clinical trials.

In the following example we will consider how information was collected to establish the safety and efficacy of a medication. Imagine you are a participant in a clinical trial of a new medication to treat bacterial sinusitis (a common bacterial infection in which the sinus passages are colonized by bacteria, often following a cold or upper respiratory infection). The study is a randomized controlled trial (RCT), and you have been randomly assigned to receive either a new treatment or a standard treatment to treat your sinusitis. You have been prescribed two tablets to be taken twice daily—one active medication, one placebo—so neither you nor your physician knows which is which (this is a double-blind, placebo-controlled trial). You return to the clinic weekly for follow-up study visits. At each visit, the research nurse asks about your adherence to the study treatment. She might use a specially designed instrument to systematically inquire about your adherence behavior. "Did you take all of your study medication since the last study visit?," she asks. As it happens, you were feeling better yesterday, and though you know you should not have done it, you did not take the whole dose—just one of the two pills you were supposed to take. Worse, you do not recall which pill you did not take, the one in bottle A or the one in bottle B. For most studies of interventions including medications, it is important to know whether all the drugs were taken, yet in the absence of biomarkers, such as drug levels, to monitor adherence, the self-report of the participant—you in this case—may be the only measure of adherence. There are powerful influences, both conscious and subconscious, that can cause someone to be less than truthful: these influences can include not remembering, fear of appearing socially undesirable, concern about being dismissed from the study, and many others. Alone or in combination, these factors may influence your response to the nurse's question. In this case, you may not recall which medication was taken.

The methods by which the question was asked also may influence your answer. Did the nurse say "Did you miss any of your doses?" or "Did you miss your last dose?" or "Did you miss any doses today?" or "How many of your doses did you miss?" or "Many people occasionally miss their study medicine; did you miss any of your doses since your last study visit?" How the study nurse asks the question, her attitude and affect as she asks the question—her tone, inflection, and body language—will influence how you respond, as will the wording she selects. As **Table 1-1** outlines, the method of inquiry by which this data point was collected becomes important with respect to the data themselves, which yield the results of this study. In addition, the method can also affect your subsequent behavior. If you give a socially desirable but inaccurate response, you may no longer wish to participate in the study, perhaps out of embarrassment. Or perhaps you remain in the study but in the future do not feel comfortable telling her the truth about your adherence behavior. Each of these outcomes will have a direct effect on the data that were collected, how we interpret the study, and how safe and effective we think the drug is.

Table 1-1 How a Question Is Asked Matters

There are many ways of asking about adherence to a medication. Some of these ways are provided in the following table. If you were asked these questions, imagine how your responses might be influenced by the inflection of the interviewer's tone, whether others were listening in on your interview, or how you feel. Examples of considerations are given in the right-hand column; more exist, of course, so when you are working with adherence survey data, be sure to consider how the method of inquiry may affect the data you collect. When possible, collecting a biomarker (such as a blood level of the drug in question) is a great addition to the study. In addition to being an ideal outcome measure on its own, even if there are not sufficient resources to measure levels on all clients, a biomarker can be used to validate data obtained through interviewing.

Types of adherence questions	Examples of methodological considerations
Did you take your last dose of the medicine?	Which medicine? Does this refer to the doses that were prescribed or the doses that the person wanted or intended to take?
When was the last time you missed taking any of your medications?	This might yield a nonspecific response such as "the other day" or "not lately" unless carefully constructed close-ended responses are provided. If this is a qualitative study, however, the question may be modified to be a more appropriate open-ended question.
How much of your medication did you take in the past week (100%, 75%, 50%, 25%, 0%)?	How does the respondent decide how doses correspond to percentages? What if there is more than one medication type?
How much of your medication did you miss in the past week (100%, 75%, 50%, 25%, 0%)?	As above. In addition, it may be easier for some respondents to think about how many doses were missed rather than how many were taken. Some respondents may be more comfortable disclosing the number of doses they took as opposed to the number they missed, despite its being more difficult to recall and/or calculate.
Did you have any trouble taking your medication?	It may be difficult to elicit truthful information asking this question, particularly if the person doing the interview is also the provider. In the event that there are barriers to adherence, it can sometimes be hard to tell one's own provider. Even if the patient can articulate his or her concerns, knowing what the barriers are in oneself can be challenging.
Did you take your medication as directed?	Does this mean on time? Right amount? Right time of day/night? With regard to food or rest? It is important for the question to be clear about exactly what is being asked.
Biomedical methods to assess adherence	Pharmacological methods used to assess drug levels. Can be highly variable depending on host characteristics, time specimen was taken, adherence to visit schedule for sampling.
Electronic means of adherence measurement	Use of electronic caps and Wisepill-type devices, which collect and transmit data each time the cap is opened. Not necessarily a true reflection of adherence, because the bottle or dispenser can be opened to retrieve several doses of the drug to be put in another carrying case and the person can open/close device to feign adherence.
Innovative new approaches	New devices becoming available that include microchips inserted into pills. When digested, they release a tiny, harmless electronic signal to indicate that the pill has actually been consumed. This may be a new gold standard, but barriers will still be in play to measure adherence, particularly for some drugs that are likely to be sold and consumed by people other than the intended participant (e.g., it may show that the pill was taken but not necessarily by the intended person).

Note how the methods of asking a critical question about adherence can alter the study as a whole on several levels. The amount of drug taken, your feelings about the study, and your decisions to stay in the study and fully disclose your actual behavior may be affected by a seemingly simple method: *how one question was asked.* All of this affects whether we end up believing the drug is effective or not.

Had you stopped taking the drug or stopped participating altogether, or if the drug were found more effective than the placebo, later analyses would attempt to evaluate patient adherence to medications and the action of the drug. But had you silently decided not to disclose your non-adherence, detecting this through any statistical or other analytic methods would be challenging; while statistical analysis can adjust for random error, it is not able to adjust for bias introduced into the study in nonrandom ways. Thus, the method interacts with not only the data but also the behavior of participants and with other elements of the study, potentially introducing bias into the study and, ultimately, into the conclusions.

Let us continue with this same example to explore additional impacts of the methods on our findings. Imagine now that you stay home from school because of your sinusitis. Midday, you receive a telephone call from a survey research firm. You have been randomly selected to participate in a case-control study exploring the association between condom use during sex and human papillomavirus (HPV), an etiologic agent in the development of cervical cancer. Case-control studies compare people with and without the disease of interest (HPV) with respect to exposure (condom use). In case-control studies, cases are identified along with suitable controls, and antecedent exposures (those that happened before the disease) are assessed among participants in both groups. The two groups are then compared. This process allows estimation of the relationship between the exposure and the disease and is especially useful in cases of rare (or relatively rare) diseases.

You are asked several screening questions, and it is determined that you are eligible to participate as a population-based control because you had an anal or cervical Pap smear in the past 12 months and have never been diagnosed with anal or cervical dysplasia (abnormal cells that may progress to cancer). You respond to a 15-minute survey with questions about your demographic attributes, routine screening, clinical and sexual behavior characteristics, and overall health and health utilization behavior.

What makes you different from someone who would not be selected as a control for this study? One salient difference between you and someone else is that you were sick and home from your usual activities at the time the interviewer called. Individuals at work or school are in general healthier than those at home during the daytime. Other differences between you and someone not selected include your having access to a phone, understanding the questions, speaking the language of the interviewer, and willingness to participate. Whatever the ultimate findings of the study, one must take into account the fact that the control participants who were selected can differ substantially from those who were eligible but not selected because of the sampling method employed. If we do not consider differences between those who were enrolled in the study and those who were not, our internal validity—how accurately our study evaluates this exposure/outcome relationship in the study population—may be limited. In addition, applying (generalizing) our findings to a larger population later may be difficult.

Thinking about methods is often like being a detective, ensuring that things are interpreted correctly, starting with how the study was conducted. If you ignore the differences between the case and control populations and look only to the information on the surface, you might draw the wrong conclusions: it could seem as if those in the control group, drawn from people who were at home during the day, are more likely to be ill and have the exposures of interest. This could be because those at home are sicker to begin with and thus more likely to have anal or cervical dysplasia, HPV, and poor condom use—for reasons having nothing to do with the research question in mind. Such a possibility prohibits drawing any conclusions about the relationship between condom use and dysplasia, as the investigators were hoping to do. **Table 1-2** displays hypothetical differences between cases, controls, and persons not selected for the study.

Now that we are about to expand on the core set of methods used in epidemiology, keep in mind that while we are seeking information, we also need to know how that information is collected, as shown in these examples. Always consider how the data were gathered: the study protocol, who collected the information, how the questions were asked, who was included in the study and how they were selected, and which data points were analyzed. Methods have a tremendous effect on the resulting data and how we will answer our research questions. Innovating and implementing creative methods to validly ascertain information is a key methodological challenge in the field.

Table 1-2 Differences among Cases, Controls, and Nonparticipants

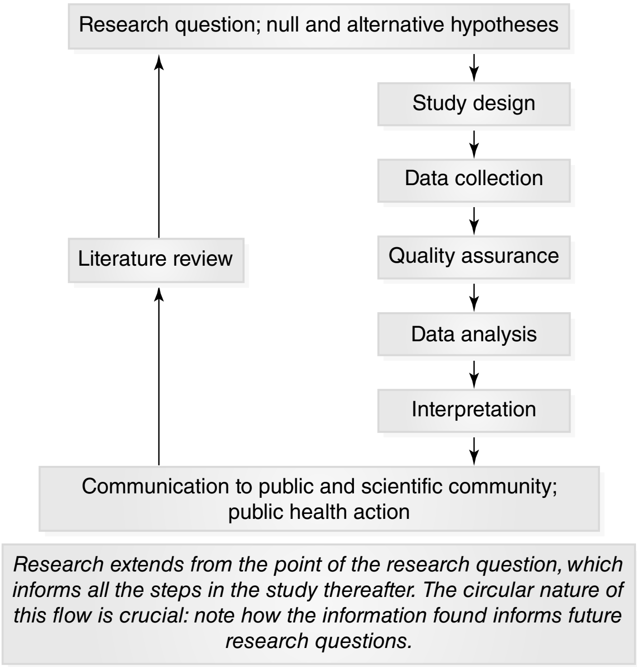
Hypothetical differences among cases, controls, and persons not selected in a case-control study of condom use, HPV, and cervical cancer are shown. These differences may not be related to the exposure (condom use) or the outcome (HPV) at all but may still affect how we are able to investigate this relationship.

Case	Population-based control	Person not selected as control
Has insurance	Sick when called and asked to participate, thus at home and not at school/work	No phone at school or job during the day
Seeks gynecological or STD-related care	Might wish to seek care, but not motivated on own	No insurance
Interested in and volunteers for research study	Interested in and volunteers for research study	Interested in research study but does not have time to participate because of work
Lives near research or clinical center	Knows about research or clinical center but has never been there	Does not live near research or clinical center
May have increased/decreased risk factors for HPV or cancer	May have increased/decreased risk factors for HPV or cancer	May have increased/decreased risk factors for HPV or cancer
May have sought care and joined study because she does not use condoms	May have sought care and joined study because she does not use condoms	May be afraid to join study because she does not use condoms

Research Question and Hypothesis Development

Before we delve more deeply into the design, implementation, and analysis of epidemiologic studies, let us review the structures central to all study designs. The first step is to identify your research question; to find an answer, one must fully articulate the question first. The research question and null hypotheses are the bases for each study's design, conduct, analysis, interpretation, and dissemination.

Your research question is derived in great part from a thorough understanding of the scientific literature in your area of inquiry. A comprehensive literature review is needed before embarking on any study and in concert with developing your research question: they contribute to each other in a reciprocal fashion. First, you identify the question of interest and then phrase that question in a form that makes it testable. Once a testable hypothesis has been generated, variables to operationalize the question are specified. Then the study is conducted to measure the exposure and outcome of interest. Once quality checks have been conducted to ensure that the data were properly collected and analyzed, information from the study is interpreted and public health action is taken, as indicated by the study. Methods and findings are documented so that they may later be shared and the study replicated. Information gained from your study will then be used as the basis for the next study, in which unanswered questions may then be addressed. This is the cycle of science, as indicated in **Figure 1-1**.


Research question; null and alternative hypotheses
Study design
Data collection
Quality assurance
Data analysis
Interpretation
Communication to public and scientific community; public health action
Literature review

Research extends from the point of the research question, which informs all the steps in the study thereafter. The circular nature of this flow is crucial: note how the information found informs future research questions.

Figure 1-1 The cycle of science.

For practice, we will develop a research question and null hypothesis and follow them through to see how they inform the study when we begin analytic approaches several chapters from now. Here is the research question: Is there an association between salt intake in excess of the recommended daily allowance (RDA) and hypertensive blood pressure? This question should be based not only on curiosity but also grounded in a literature search. The question you ask should fill a gap in the literature whenever possible and be based on concepts that previous studies have shown to be of interest to explore.

Once you have identified your research question of interest, it can be used to define your null hypothesis (H_0). The H_0 is the statement of no difference and provides the basis for statistical testing, which assesses the role of chance in the study findings. Its counterpart is the alternative hypothesis (H_A), which expresses the opposite of the null hypothesis and characterizes what we expect to find as the result of our study. The research question and null (and alternative) hypothesis are not the same: the research question is the relationship you are investigating, whereas the null hypothesis provides the comparative basis for your investigation. In many respects, the strength of your method lies in the articulation of your null hypothesis and how you operationalize it. The benefit of stating the question as a null hypothesis is that it informs our thinking about what to measure and how to decide that an association exists, ultimately shaping our study design. Our null hypothesis could have many forms:

- H_0: Blood pressure (BP) among people who consume salt $>$100% of the recommended daily allowance (RDA) *is the same* as BP among people who consume salt $\leq$100% of the RDA.
- H_0: BP among people who consume excess salt *is the same* as BP among people who do not consume excess salt.
- H_0: Each subject's BP over 2 weeks of consuming daily salt $>$100% of the RDA *is the same* as the BP of the same people over 2 weeks of consuming daily salt $\leq$100% of the RDA.

Moving from the basic level to the intermediate level, the next steps are to learn how these research questions are applied in terms of data and analysis. Here is an example of what we might see in terms of the data themselves as we apply this research question.

For this example, our primary independent variable (X) of interest is salt consumption (whether in excess of 100% of the RDA), and our dependent variable (Y) is BP. The research question and background literature on the topic will define how we construct our variables, something the research question above did not address. There are multiple ways that these variables could be defined, or operationalized, that is, how we can concretize the research question. We could measure both X and Y as categorical (categorized based on the median, or the tertile, or the quintile, perhaps) or as continuous. We could base them on a clinically relevant cut point or on data from a prior screening test that indicated a particular value was associated with the outcome of interest. (Note: this is where we must have consulted the existing literature; without doing so, we will not be able to collect or examine the data in any meaningful fashion.) Similarly, we must propose a means to define how we examine the dependent variable—BP. Categorically? Continuously? According to the guidelines for hypertension? What is the cut point for hypertension? Based on a data-derived definition? All of these criteria must be defined in advance and proposed as a part of the study. Often it is the definition of the variable itself that allows a study to fill an important gap in the science. We will continue with this example in a later chapter.

Research questions usually can be addressed in more than one way. The question in this example could be addressed with an experimental design that randomly assigns patients to interventions that would modify excess salt intake RDA; or in a cohort study, in which patients who are high-salt consumers are compared with low-salt consumers; or in a pre-/posttest study in which participants' blood pressures are measured successively during high- and low-salt diets. Each of these designs has unique strengths and limitations, and would each answer a different research question. Each design also has its unique methodological challenges, which need to be overcome. Irrespective of the design approach taken, the research question defines the association of interest, and the null and alternative hypotheses provide specific detail of how the relationship will be evaluated. Null hypotheses may be two sided as previously, or they may be one sided, with one-sided alternatives. These are identical to their two-sided counterparts except that they specify a direction. For this example, a one-sided null hypothesis might look like this:

- H_0: Blood pressure (BP) among people who consume salt in excess of the recommended daily allowance (RDA) $\leq$ in excess BP among people who consume salt less than or equal to the RDA.
- H_A: BP among people who consume salt in excess of the RDA is greater than the BP among people who consume salt less than or equal to the RDA.

Note that in a one-sided null hypothesis, the equal sign remains in the statement of the null but not in the alternative. Using a one- versus two-sided null hypothesis affects how we treat the significance levels.

Operationalizing or Defining the Variables

The selection of variables to evaluate is an important decision. Rather than having vague independent and dependent variables, all types of designs in epidemiologic research, from descriptive to observational to experimental, require clearly identified variables. For example, imagine an experimental design evaluating a campaign to reduce exposure to mosquitoes during a West Nile virus outbreak. Consider for a moment the ways that exposure to the campaign might be measured: number of people living near a social marketing campaign billboard; driving near the billboard; reporting in a survey that they saw the billboard, read the billboard, and took action based on the billboard, and so forth.

Operationalizing the variables of interest is a necessary step that needs to be done well. Without explicit operationalization, it would be easy to end up with information that is invalid. Consider all the ways that the outcome "exposure to mosquitos" might be measured: purchase of DEET-based insect repellant, number of people coming to the emergency room fearing exposure, number of reported and confirmed cases of West Nile virus, number of insect bites on the arms and legs of study participants, and so forth. Based on previous research and the research question at hand, the investigator must be explicit about how the variables under study will be defined. These need to be selected in a multidisciplinary fashion, as well, to ensure that no facet of the relationship or its study is being missed. We have not yet reviewed confounders or effect modifiers, but considering them at the inception of the study is necessary. If one gathers data at the beginning on potential confounders and effect modifiers, their effect can be analyzed. It will become nearly impossible to ascertain data on such variables after the study is over, as people move, environments change, and most people may not remember what was taking place in their lives at the time of the original data collection.

It is imperative that we consider this in the beginning, because if we neglect to, it is possible that after the study data are collected, we might think "if only we had these data in another form." For example, imagine we are conducting a study of lung cancer and an environmental toxin that has recently been discovered. We will obviously wish to include smoking behavior (as well as exposure to secondhand smoke) in our data collection, as it may well be a confounder or effect modifier. It is not enough to consider potential concern *smoking*; we need to think systematically about how we will collect data about it and whether it will meet our analytic needs. If we are looking at an existing data set for secondary analysis, we may have smoking as a binary variable (yes/no) but not in terms of cigarettes per day per year or depth of inhalation or other variables that we may need. Thinking it through from the start, and working from a firm knowledge of the previous research in the area, will allow us to search for the appropriate data set rather than lose time and resources trying to work with an inadequate one. Studies do not happen by accident, and building in processes by which we can think about data needs *a priori* is key. In the event data are not available at all, at the very least, we will be aware of and can speak to the limitations in our project and quantify and qualify them to the extent possible.

Here is a slightly more complicated example. Let's consider the various ways of operationalizing the variables associated with the research question "Is substance abuse status associated with unmet needs and HIV-related immunocompromise?" We need to operationalize the following variables:

- Substance abuse
- Unmet needs
- HIV-related immunocompromise

For this example, imagine working at an injection drug use (IDU) clinic and wanting to assess the relationship between substance abuse and needs. Do we want to ask participants about substance abuse and unmet needs that were occurring at the start of the study? Within a certain time frame of the study (i.e., within 3 months of enrollment)? At the moment of the assessment? Do we want to allow these variables to change over time with the status of the participant (also known as time-dependent covariates)? All of this information must be clearly specified for the definitions to be systematically applied to all participants by all research staff. Otherwise, the study will be conducted differently depending on the participant, the day, and the interpretation of the researcher collecting data. When we consider eligibility criteria for the study, this is especially important. A study that excluded current IDUs but allowed former injection drug users who had not used drugs within the past year would yield a very different research population than one that excluded anyone who had *ever* injected drugs. How might these variables be operationalized?

Substance abuse (independent variable):

- History of ever using illegal drugs as assessed by study-specific drug inventory at baseline
- Current (within 3 months of study enrollment) use of illegal drugs as assessed by study-specific drug inventory at baseline
- Current (within 3 months of study enrollment) use of any substance as assessed by study-specific drug inventory at baseline, including tobacco, cigarettes, and prescription medications
- Failure to remain drug free for 3 or more weeks as determined by substance abuse counselor
- Use of validated measures

For each of these variables the factors within them will also have to be defined. How will use and timing of use be measured—by self-report, drug levels, observation of track marks, or other measures of drug use? Each element of each component of the variable must be clarified and operationalized.

Unmet needs (dependent variable):

- Self-report of unmet needs at baseline in any one of the following domains using a specific instrument:
    - Activities of daily living
    - Social support
    - Family support
    - Healthcare service utilization
    - An index of 10 or greater on an acuity scale
- Self-report on questionnaire of having lived in a homeless shelter, been in a transient living situation, had a hospitalization, or engaged in survival sex or commercial sex work at any time during the previous 12 weeks
- Answering a question at baseline—"Do you have any needs that are not currently being met by your ancillary service care program at this time?"—as yes or maybe

HIV-related immunocompromise (covariate, or confounder, or effect modifier)

- Self-report of a CD4 cell count <200 cells/mm3 at baseline
- Study-specific detection of a drop in CD4 count of 100 cells/mm3 or greater between any two visits of more than 12 weeks apart
- Having or developing an AIDS-defining diagnosis over the duration of the study follow-up period as detected by medical record extraction

As these variables are operationalized and our protocol is being developed, it is helpful to keep track of our information in a concise table or research question matrix (see **Table 1-3**). The research question matrix is a tool to help keep our study design in order. This matrix may be modified as needed to incorporate our study's specific research questions, hypotheses, and variables. By keeping all of our research questions, exposures, and outcomes in order, we can be assured that we will not forget to measure anything or lose sight of the null hypothesis guiding our study.

Table 1-3 The Research Question Matrix

Research question	Dependent variable(s) (outcomes)	Independent variable(s) (exposures)	Potential confounders	Potential effect modifiers
Example: Is substance abuse status associated with unmet needs and HIV-related immunocompromise?	Unmet needs (self-report on baseline questionnaire; self-report on questionnaire of specific experiences in previous 12 weeks; or yes/maybe response to unmet need question at baseline)	Substance abuse (history of illegal drug use reported on inventory at baseline; illegal drug use within 3 months prior to baseline drug inventory; use of any controlled substance within 3 months prior to baseline drug inventory; or failure to remain drug free for ≥3 weeks as assessed by substance abuse counselor)	HIV-related immunocompromise (self-report of a CD4 cell count <200 cells/mm^3 at baseline; study-specific detection of a drop in CD4 count ≥100 cells/mm^3 between any two visits of more than 12 weeks apart; or having/developing an AIDS-defining diagnosis over the duration of the study follow-up period as detected by medical record extraction)	HIV-related immunocompromise (See Potential confounders.)
You can also add in null and alternative hypotheses to keep together	What is your outcome of interest? How are you operationalizing (defining) it?	What exposures are you assessing in conjunction with the independent variable to the left? How will each be operationalized?	What independent variables should you consider as potentially confounding your understanding of the relationship between the dependent and independent variables? Think them through up front so you do not forget to collect data on those that you need to investigate.	What independent variables should you consider as potentially altering the nature of the relationship at different levels of the independent variables? Just as with confounders, think them through up front so you do not forget to collect data on those that you need to investigate.

Considering Causality

To some degree, our understanding of causality in nature is primal. Driven more by reflex than insight, we know from our earliest moments that if we touch a hot stove we will be burned. Causality pertains to study designs, interpretation of data, and communication of results, but here, we will introduce the topic from a conceptual viewpoint as a foundation for an examination of epidemiologic methods.

As in the previous examples of research questions, we want to understand a relationship between an exposure and a disease, and through understanding this relationship, identify ways to reduce the incidence and prevalence of disease either by primary or secondary prevention or treatment. While occasionally we might explore a relationship just out of scientific curiosity, in general, we are trying to understand public health phenomena so that we can improve the health of populations. There is a logical order that we assume exists and that provides the underpinning of our form of study: increased salt intake *leads to* a health outcome of interest. Cell phone use *leads to* a health outcome of interest. Use of a medication *leads to* improvement in care or as the case may be, adverse events. This order allows us to consider ways that we could improve public health: decrease salt intake, change cell phone use behavior or technology, or alter (increase, decrease) use of medications to improve health outcomes and minimize risks. What we are looking for is temporality of effects, A precedes B, which is a necessary though not sufficient piece of demonstrating a causal relationship, A causes B. Our research questions strive to inform the question of causality so that we can contribute to the understanding of the phenomena affecting public health.

One limitation to our being able to infer temporality and, in turn, causality, stems from the fact that in many observational designs, we cannot control exposures and so cannot validly assess temporality, preventing us from inferring causality. It is critical that as epidemiologists we are able to distinguish between study designs that have the capacity to demonstrate temporality or causality and those that do not. We have the luxury in some cases of being able to employ experimental designs, whereby we can randomly assign exposures and follow for outcomes, solving the issue of temporality and causality. We also have the ability to construct cohort studies or leverage existing records so that we can establish temporal relationships by identifying exposures that precede disease.

Often, however, we are unable to utilize these designs and must instead rely on ecological, cross-sectional, case-control, or retrospective cohort studies that allow us to evaluate exposures after the outcome has occurred. These designs are powerful and can inform us about relationships between potential exposures and outcomes, but are complicated by the reality that the data on the exposures may not have been collected before the outcomes occurred. Even when we can benefit from experimental or prospective cohort designs, we often encounter challenges, such as nonresponse, missing records, loss to follow up, and other biases, that limit our ability to establish causality with the certainty that we might prefer. For this reason we need to be especially careful when designing our studies, analyzing them, and communicating our findings that we do not overstate the role causality plays in them. As you become more familiar with the methods and each of their unique strengths and limitations. The more you will be aware of the nuances of each of the methods, and the more naturally and accurately you will be able to communicate your findings. You will become increasingly facile using phrases such as "We found an *association* between X and Y" rather than "X *causes* Y." You may already by this time be familiar with Bradford Hill's classic essay on causal evidence; it is an excellent resource to help consider the basic tenets of causality.

Bradford Hill Criteria for Causality

Austin Bradford Hill (1897–1991) proposed what is still used today—a thoughtful set of criteria for causality. One powerful feature of these criteria is that they underscore the importance of methods and contributing to the scientific dialogue through writing. They look at a body of work—not just one study on its own. As a refresher, refer to **Table 1-4**.

Table 1-4 Bradford Hill Criteria for Causation

Criterion	Refers to
Strength of association	"The stronger the association, the less likely that it is due to confounding or a non-causal relationship." Yet the flip side is "We must not be too ready to dismiss a cause-and-effect hypothesis merely on the grounds that the observed association appears to be slight." (Note: In the original paper, Hill also adds important points about the difference in searching for etiologies vs. having practical importance and the difference between assessing absolute rate differences and rate ratios. In addition, the utility for this criterion alone—although Hill's point is that not one of these really should be taken on its own—depends on methods that validly collect data. If bias is introduced or the design is highly flawed, no association, strong or weak, will be able to tell us much about causality or the relationship.)
Consistency	"Has it been repeatedly observed by different persons, in different places, circumstances and times?"
Specificity	"If, as here, the association is limited to specific workers and to particular sites and types of disease and there is no association between the work and other modes of dying, then clearly that is a strong argument in favour of causation."
Temporality	"Which is the cart and which the horse?"
Biological gradient	"dose-response curve"
Plausibility	"It will be helpful if the causation we suspect is biologically plausible. But this is a feature I am convinced we cannot demand. What is biologically plausible depends upon the biological knowledge of the day." (This latter phrase is brilliant and notes that much of what we think is plausible is determined by our abilities and technology—it is hardly absolute.)
Coherence	"[T]he cause-and-effect interpretation of our data should not seriously conflict with the generally known facts of the natural history and biology of the disease."
Experiment	"Occasionally it is possible to appeal to experimental, or semi-experimental, evidence. For example, because of an observed association some preventive action is taken."
Analogy	"In some circumstances it would be fair to judge by analogy. With the effects of thalidomide and rubella before us we would surely be ready to accept slighter but similar evidence with another drug or another viral disease in pregnancy."

Reproduced from Bradford Hill, A. The Environment and Disease: Association or Causation?, *Journal of the Royal Society of Medicine* 1965 May; 58(5): 295–300.

Considering the Importance of the Counterfactual

In descriptive designs, we are able to use many tools from our toolkit to describe data: plotting, tabulating, measures of central tendency and dispersion, graphing, mapping, and so forth. These powerful tools provide valuable information to us and are used in every analysis prior to bivariable or multivariable techniques. Descriptive designs do not require a comparison group; we just describe the data we have. Often, that description will visually or intuitively suggest relationships to us, yielding hypotheses that can be quantified using analytic designs.

Many analytic methods rely on comparison groups. As we move from describing information to assessing null hypotheses, we usually will need to establish a comparator, a reference point. It is not enough to say that "Children in group B are tall" because to be "tall" we must know "Compared to whom?" Group B children might be taller than those in group A but shorter than those in group C. Use of the referent is one feature that distinguishes between descriptive and analytic designs.

One unique characteristic of epidemiologic methods is that they maximize use of the counterfactual. Not only do we often require a comparator, that comparator is nearly always a counterfactually defined event. For example, the patient either has cervical cancer in situ or does not have cervical cancer in situ. The outcome of cancer can be studied, or if preferred by the investigator, the outcome of *absence of* cancer can be studied. We can code the condition of cancer (operationalized as appropriate for the study, for example, based on pathology reports) = 1 or the *absence of* cancer = 1, depending on the circumstances. If we hypothesize that a newly identified sexually transmitted infection (STI) is associated with increased incidence in cancer, it may be that the former would make more sense (e.g., having this STI is associated with increased cancer incidence).

On the other hand, if we are hypothesizing that a new drug is associated with reductions in cancer, the latter might be more intuitive (e.g., taking the experimental drug is associated with decreases in cancer incidence). If we hypothesize that an exposure is associated with increased incidence of an outcome, we may want to code the outcome as 1; if the exposure may be associated with reduced incidence of the outcome, we would code the outcome as 0 and its absence as 1. Central to using the counterfactual in this sense is that each condition must be mutually exclusive and comprehensive: all participants must be definable as either having or not having the outcome of interest. There can be no participants that simultaneously have and do not have the outcome. There can be no participants who "maybe" have the outcome. In this way, if all the patients in the sample = 100%, then (cases ÷ 100) + (controls ÷ 100) = 100%. We can define the measures variously, in concert with our research question and null hypotheses, but we must use non-overlapping, mutually exclusive categories in all cases.

In **Figure 1-2**, imagine that we are looking at an RCT of a drug to prevent cancer progression among women. Here the analysts can code, in conjunction with their hypothesis, that the drug will prevent development of stage IV cancer among participants. This will mean that they phrase their results in the context of prevention, not disease. But in **Figure 1-3**, now imagine that we are doing an observational study of drinking alcohol and its association with progression to stage IV cancer among a cohort of women. Here the outcome might be coded in reverse: progression

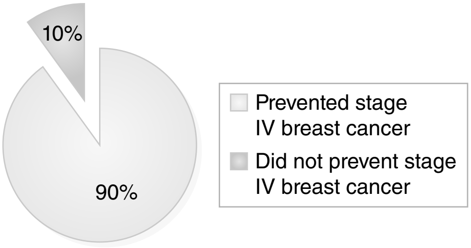


Figure 1-2 Mutual exclusivity and the counterfactual, part I.

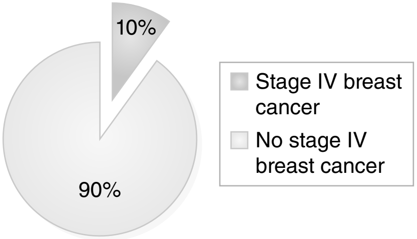


Figure 1-3 Mutual exclusivity and the counterfactual, part II.

could be coded as 1 and nonprogression as 0. Here you can see that there is no one right answer; both can work if they are mutually exclusive and carefully defined. Even within each study based on the findings it may well be easier to talk about either outcome—it just depends on what is being modeled. As long as the codes are consistently referred to in terms of what was modeled in the analysis and the outcomes are mutually exclusive, we can look at either; that is the beauty of the counterfactual. (Just remember that with an RCT, provided it was analyzed with intention to treat analysis and requisite assumptions met, the authors could reasonably say that the medication *prevented* progression. In the cohort study, this will not be possible; we can measure only associations in observations studies.)

Through provision of a counterfactual, we automatically have a referent. "Compared to patients who did not have cancer, those who did were more likely to have taken the drug of interest." If we coded absence of cancer as 1, then this statement would yield an odds ratio (OR) >1.0. We could also have coded cancer as 1, in which case the statement "Compared to patients who did not have cancer, those who did were more likely to have taken the drug of interest" would have yielded an OR of <1.0. If we can compare groups to one another and make use of this referent, we go from describing to analyzing, from characterizing to comparing. Then we can say that groups are different from one another—more than, less than, taller than, heavier than, and so forth—and begin our business of describing relationships between variables.

The approach using the counterfactual can easily be used for categorization with any variable. But it also can be used when treating continuous variables, with a little bit of data management. If we were considering titer levels (of an antibody or chemical, for instance) originally collected as a continuous variable, using them categorically is as simple as developing a cut point; each participant is either in the high category or the low category based on the titer cut point. It is necessary to be sure of a few things when using multiple categories. First, the whole range of numbers must be accounted for (for example, >10 vs. ≤ 10 or ≥ 10 vs. <10, but not >10 vs. <10). Additionally, when more than one category is established, the referent must be clearly designated (group 1 as referent, with group 2 compared to group 1 and group 3 compared to group 1). The counterfactual is not present when we look at the three groups individually, but it is in each of the component comparisons (group 1 vs. not group 1, etc.). Finally, realize that reducing continuous data to categorical data has a cost in precision, which at times may result in substantial loss of understanding or residual confounding. Use of the referent condition in our methods is one of the chief strengths of the tool kit.

Multifactorial Causation

Understanding the role causality plays in relationships between independent variables and dependent variables is critical to public health. To a large extent, our ability to improve the health of populations depends on our ability to understand causal relationships: to improve public health we need to identify threats public health it and ways to eliminate them. Similarly, we strive to understand how treatments effect cure and how system improvements allow better access. Implicit in these goals is the concept of causality: if the exposure did not cause the disease, prevention of exposure will not prevent disease. If the treatment did not cure the disease, its provision will not improve public health. Thus the supposition of causality across time and studies—even though it cannot be shown in each study—is key to the ability of information to transform into public health action.

Causality is a complicated concept. It would be convenient if every outcome of interest had only one definitive cause, yet this is seldom the case. More common is the concept of multifactorial causality: the presence or absence of factors contributes to an outcome, a constellation of phenomena that together lead to an event or condition. We are going to look at this concept in two ways: conceptually and symbolically.

Conceptualization of Multifactorial Causation

As a framework for discussing causation, let's examine an example in which there is a 7-year-old boy swimming in the ocean. His safety is dependent on many characteristics of the day, including but not limited to:

- Characteristics of the child: How well he can swim, how strong he is, how accustomed to swimming in the ocean he is, how calm he is, how well he breathes (asthma? illness?), how far he is from the shore. Does he panic when confronted with trouble getting back to the shore?
- Characteristics of the water: The nature of the waves, riptide, current, the temperature of the water

- Characteristics of those around: Are adults (lifeguard, parents, others) nearby and are they paying attention? Do the adults supervising him know how to detect a struggling swimmer? Do they know how to rescue him? Are they healthy enough to perform a rescue? Do they panic?
- Other characteristics of the environment: Is it a busy beach? Are there other barriers or facilitators to swimming safely to the shore?

In **Table 1-5** we are going to look at the relative addition or removal of characteristics in association with his drowning. Note that you could look at these same factors in association with the counterfactual outcome, that of his not drowning, as well, in which case we would be taking a prevention approach.

We can see that even in this hypothetical example with artificially few variables in play (and falsely dichotomized variables at that), many factors come into play in terms of the outcome we are studying. In actuality, there are many gradations of each of the variables and many additional variables that would be added, but this information highlights that the addition or removal of characteristics changes the estimate of risk of the outcome. Just because the child is a poor swimmer does not in isolation mean that he will not make it back safely to the beach.

Rothman and Greenland (1998) and other authors have provided a helpful framing of the multifactorial causation question using a sufficient-cause model. They refer to sufficient cause as a minimal set of conditions that is sufficient for the outcome to occur. An outcome of not drowning, in this example, could be the result of having an alert lifeguard noticing the child struggling in the water. Thus even in the presence of dangerous waves and a compromised child, this could be sufficient to *prevent* drowning. Conversely, having a child who is compromised and with a rough sea and a parent who cannot swim supervising the situation could be sufficient to *cause* drowning. There is also the constellation of unknown factors associated with the child, environment, and so forth. There are numerous sets of sufficient causes that can lead to any given outcome.

Table 1-5 Multifactorial Causation Examples

Characteristics of child	Characteristics of ocean	Characteristics of supervision	Outcome is
Strong swimmer	Riptide, strong current	Poor	Not drowning
Weak swimmer, compromised health status	Riptide, strong current	Poor	Drowning
Weak swimmer, compromised health status	Calm ocean	Good	Not drowning
Weak swimmer, compromised health status	Riptide, strong current	Good	Not drowning
Strong swimmer	Riptide, strong current	Good	Not drowning
Strong swimmer	Calm ocean	Poor	Not drowning
Weak swimmer, compromised health status	Calm ocean	Poor	Possible drowning

As part of the sufficient-cause model, let's now consider each of the factors contributing to an outcome (child, sea, adult, unknown factors) as a component cause (pieces of the whole that make up the sufficient-cause mechanism). A necessary cause is a component factor that is a part of each and every sufficient-cause mechanism. These are the things that are central to the outcome's occurring, such as the child must be in the ocean to drown. In the case of risks for, say, uterine cancer, the person must have a uterus. These are often relatively obvious factors, but they are important to consider when developing the proper denominator of interest (e.g., who is at risk) and the proper population to generalize findings to. We can illustrate this relationship using a directed acyclic graph (DAG) (**Figure 1-4**) to help sort out the causal relationships and contributory factors to the outcome, as well as to graphically and conceptually account for the issue of confounding. In using these tools, it is important to remember that their ability to facilitate understanding is the key: we cannot look at these highly interwoven and complex concepts solely in their reduced form.

Using the diagram we can delineate the relationships and proposed causal pathways. These diagrams can be applied to any number of situations to describe conceptually the relationship between the variables and the causal directionality. They can also be applied to **Figure 1-5**.

Here is another example, one with a well-known health outcome, whereby we can include the potential for unknown factors: a Sample DAG to describe the relationship between IDU (injection drug use) and hepatitis C/HIV coinfection may be found in **Figure 1-6**.

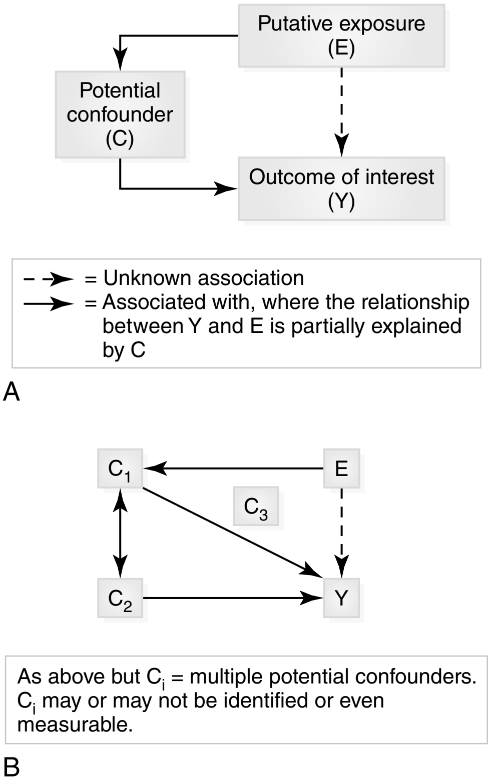

Putative exposure (E)
Potential confounder (C)
Outcome of interest (Y)
- - ➤ = Unknown association
——➤ = Associated with, where the relationship between Y and E is partially explained by C
A
C_1
E
C_3
C_2
Y
As above but C_i = multiple potential confounders. C_i may or may not be identified or even measurable.
B

Figure 1-4 Directed Acyclic Graphs (DAG).

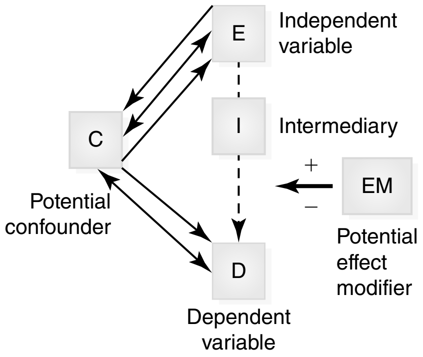


Figure 1-5 Directed Acyclic Graph (DAG).

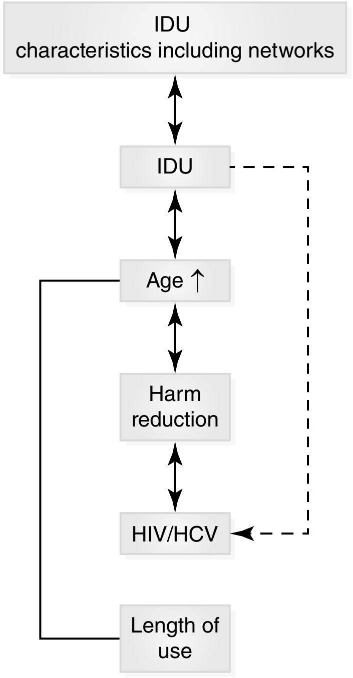


Figure 1-6 Sample DAG to describe the relationship between IDU (injection drug use) and Hepatitis C/HIV Coinfection.

When we consider epidemiologic questions, the concept of multifactorial causation is critical. Causality is seldom direct and easy to ascertain. Add in confounders and effect modifiers, and these relationships are anything but straightforward. Except for rare cases, diseases are caused by multiple factors and prevented by multiple factors as well. Looking at causality in an overly simplified way does a disservice to our field and also can undermine how we communicate epidemiological findings to the public. Just because we draw a line suggesting causal relationships does not mean there is one and does not mean there are no other ones. Being mindful of the complexity of the relationships between exposures and outcomes will allow us to further expand our epidemiologic toolkit and better understand the nuances we will encounter when applying it.

Discussion Questions

1. Write down five types of data collection that illustrate the concept of how asking a question may influence the answers given.
2. Now imagine a data collection scenario based on one of your five challenging data collection issues. Using **Table 1-6** as a template and Table 1-2 as an example, complete the table to show what might be differences between cases, controls, and nonparticipants for your hypothetical study.

Table 1-6 Differences between Cases, Controls, and Nonparticipants

Research question:

Case	Population-based control	Person not selected as control

3. Develop a research question that you are interested in considering. You might want to choose one that will correspond to a literature review or project you are planning on for your master's in public health degree. In addition to your research question, develop your null and alternative hypotheses.
   Research question of interest: ____

   ____

H_0: ____

____

H_A: ____

____

4. Consider the research question that you developed in Table 1-6. How will you operationalize (define) your variables? You will be including them in the matrix below, but thinking about your key variables now *and* operationalizing them up front will help enormously before you track them in summary form later on.
   Key outcome measure (dependent variable): ______________________

   ____

   Primary independent variable(s) of interest: ______________________

   ____

   Potential confounder(s): ______________________________

   ____

   Potential interaction(s)/effect modifier(s): ______________________

   ____

5. Complete the following research question matrix (**Table 1-7**) for your research question. Be sure to carefully consider what characteristics you will need to collect data on to answer your question. If you choose a topic that is one you will be writing on for a real project, then consult the literature to determine the appropriate variables to include.

Table 1-7 Research Question Matrix

Research question	Dependent variable(s) (outcomes)	Independent variable(s) (exposures)	Potential confounders	Potential effect modifiers

6. For each of the following null hypotheses, provide the counterfactual, and indicate the expected estimate of risk or odds based on the proposed H_0:

H_0	**Coding scheme**	**Counterfactual coding scheme (change for only the dependent variable [DV] or independent variable [IDV] in this table, though you could of course do both)**	**Expected estimate under the counterfactual coding scheme if the H_0 is rejected (OR = odds ratio)**
Example: Proportion of babies born with neural tube defects to women who consume the RDA of folic acid during pregnancy	IDV: Mother consumed folic acid = 1 Mother did not consume folic acid = 0	IDV: Mother consumed folic acid = 0 Mother did not consume folic acid = 1	OR > 1.0
Example: Proportion of babies born with neural tube defects to women who consume the RDA of folic acid during pregnancy	IDV: Mother consumed folic acid = 0 Mother did not consume folic acid = 1	IDV: Mother consumed folic acid = 1 Mother did not consume folic acid = 0	OR < 1.0

7. For the research question of interest you have chosen, create a DAG to reflect the putative underlying relationships. What does this DAG *not* reflect?