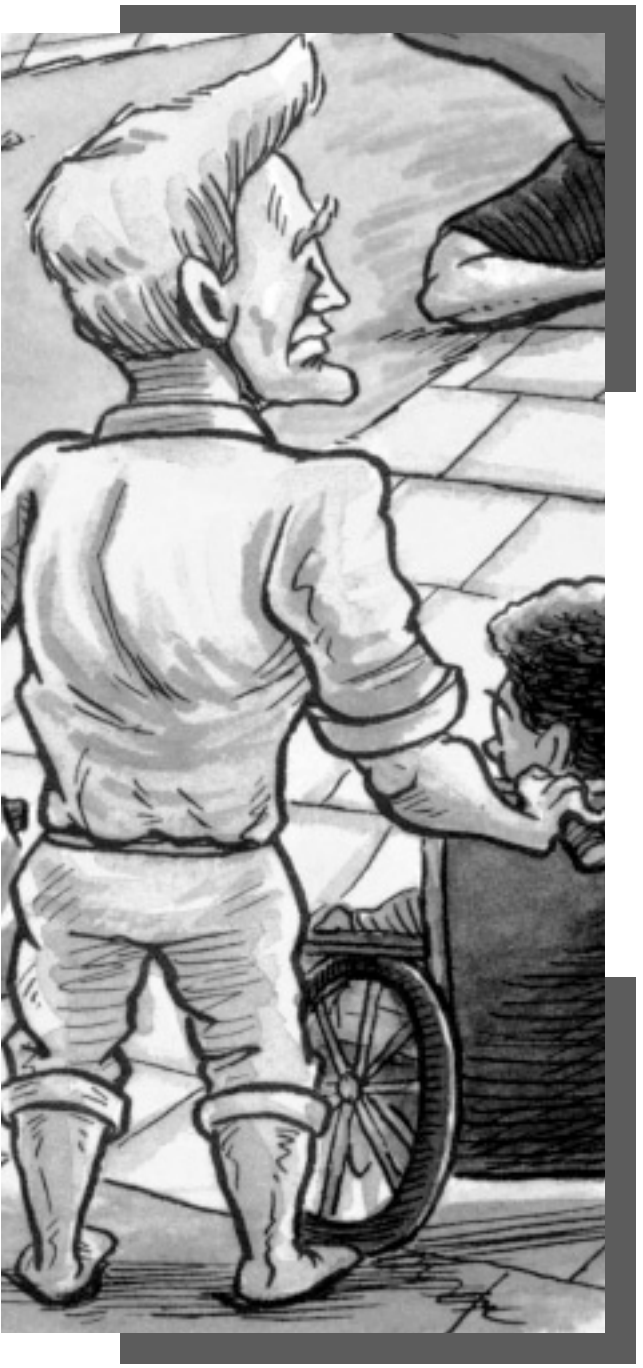
Epidemiology, Policy in Public Health, and Human Judgement



by Donald Austin, MD, MPH

The environment in which the Monarch butterfly evolved gave it a distinctive pattern of coloring. A young oriole foolish enough to catch and eat a monarch butterfly begins retching within minutes. From that time forward, that bird avoids insects that utilize the monarch's pattern and coloring.

By this experience the bird learns a useful association between a true cause and its effect. Learning this relationship is critical to the well-being of the oriole. However, because of it, the oriole also avoids similarly patterned species that provide harmless nourishment.

Evolution has also endowed humans with an internal neural system that associates specific experiences or situations with good or bad outcomes. At the very basis of our behavior is the need to identify relationships and the need to know, when encountering a specific situation, what is likely to happen next. Prehistoric man certainly survived longer after he recognized that caves surrounded by animal bones were likely to contain a dangerous carnivore. Modern humans who have experienced food poisoning and

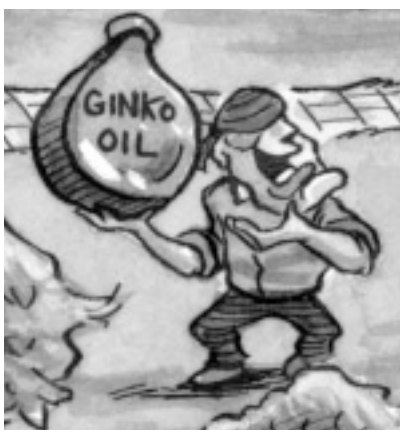
fallen ill shortly after consuming seafood paella may develop an aversion to that dish, even if the actual cause of the food poisoning was something eaten 24 hours earlier, such as alfalfa sprouts.

Modern humans in the 17th century developed a process labeled the "scientific method" to help sort fact from fiction. One of the most important applications of scientific method is to distinguish cause-and-effect associations from correlations between events that are related for some other reason. In epidemiology we would say that causal association made by the oriole between the monarch butterfly's toxicity and its distinctive pattern of coloring is not valid. Scientific method depends on the refutability of a stated hypothesis for its usefulness. For example, if one hypothesizes that the direction faced is east and the sun rises to the rear, the hypothesis is refuted. *(Please see the glossary for a definition of the scientific method as well as other terms used in the article. -ed.)*

If an association passes a series of tests and meets certain criteria, the final result is a conclusion. It is not a proof, since there is no incontrovertible positive proof in the scientific method, only sufficiently convincing evidence leading to the position that any other conclusion would be foolish. Evaluation of what is foolish and what is not depends a great deal on human judgment and the conscious and unconscious agendas of those making important decisions. Even barring their inherent scientific complexity, addressing public health problems requires far more discernment than determining from which point of the compass the sun rises.

Epidemiology

Epidemiology is an observational science. Most of analytic epidemiology is based on the concept that we are all in one big, complex, natural experiment. The strategy of epidemiology is to use different study designs and analyses to figure out who should be in "experimental" groups (cases—those with a disease) and "control" groups (those who do not have a disease), and to draw conclusions from that comparison. However, participants in an observational study are not randomly assigned to treatment or placebo groups. In epidemiology, a study's designers must create a surrogate for randomization through careful selection of participants, matching selected characteristics, adjusting for certain differences, and conducting a data analysis that minimizes the effects of not having the non-random characteristics of a true experiment. One goal of an observational study is to create the analytic facsimile of a flawless experimental study.



However, even a conclusion drawn from a flawlessly done study

Obesity, Paradigms, and Facts

It is hard to know when a prevailing paradigm may be shifting. In public health, this type of change happens slowly. We are probably experiencing more than one instance today. One candidate for a shift is the paradigm governing the dietary guidelines created by the government.

For over half a century, indirect studies have linked a high fat diet to heart disease. Likewise, high cholesterol levels are a risk factor for heart disease. Some research indicates that cholesterol levels can be altered (at least modestly) by decreases in lipid and cholesterol intake. In the late 1990's, the National Institutes of Health (NIH) and other federal and national organizations translated these findings into official dietary guidelines for the nation. Restriction of dietary fat and a major reliance on carbohydrates for caloric intake figured prominently in those recommendations. Fruits and vegetables are also prominent, with a recommendation for five servings a day.

Some scientists now believe that dietary recommendations to avoid fat and focus on carbohydrates have contributed to the epidemic of obesity in the U.S. over the last several decades. At least three long-term observational studies, the Nurses Study I and II and the Health Professional Study, have all failed to demonstrate a relationship between dietary fat and obesity. As many as five intervention studies, among them the NIH-funded Multiple Risk Factor Intervention Trial (which included interventions for smoking, hypertension, diet, cholesterol level, and exercise) failed to show that dietary fat was related to heart disease. In fact, there is no strong scientific evidence that dietary fat in a diet of appropriate calories leads to either obesity or heart disease in persons not already predisposed to heart disease.

Changing dietary recommendations flies in the face of what we have been taught for half a century. The existing paradigm regarding the causes of obesity, diabetes, and heart disease has made it extremely difficult to fund and conduct dietary studies challenging the existing low-fat recommendations by NIH and the USDA.

is not a proof, since there are no incontrovertible positive proofs in the scientific method, only sufficiently convincing evidence. Acceptance of an hypothesis as fact is often a gradual process and depends on repeat testing.

Human Judgment

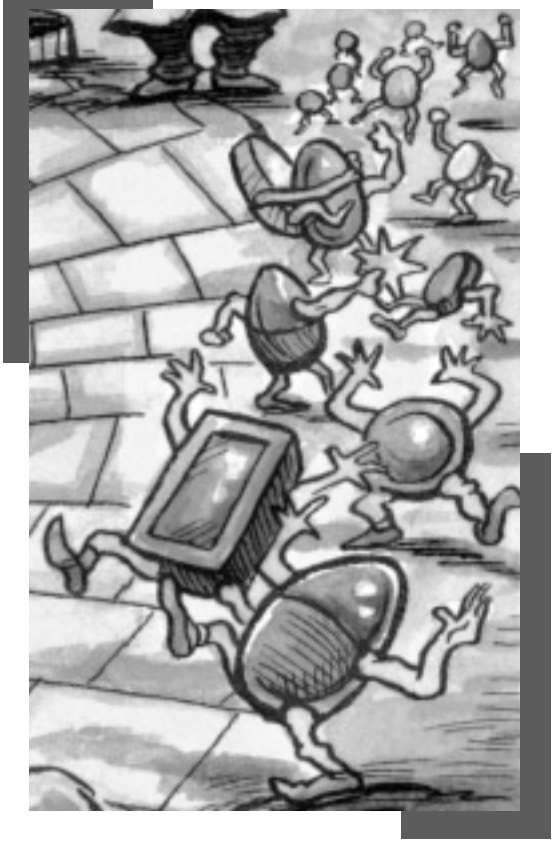
Just as the oriole learns to attribute the monarch's toxic effect to all insects with patterns and coloring similar to the monarch's, scientists also make associative mistakes. Because one person's foolishness is another's rational logic, there is room for a lack of consensus of the validity of many conclusions. The use of scientific method is subject to error because it depends on human judgment at several points. For example, a scientist must first decide on which hypotheses to test; one seldom tests hypotheses that are considered too silly to test. Then a scientist must decide on how many times and what different ways to test the same hypothesis before accepting it as valid. Both of these depend on the believability of a hypothesis. *(Please see sidebars Obesity, Paradigms, and Facts as well as When is a Fact a Fact.)*

How Findings Become Accepted Facts

Consider a situation where you have taken a bus to a large city to take care of an important task that has kept you inside an unfamiliar building until after dusk. It is now nearly time to catch your bus. Imagine that you leave the empty building by a different door than the one you entered. Outside, you are confused about which way to walk to get to your bus stop. Across the street is a small child on a skateboard. You cross the street and ask directions to the bus route. The child points and answers, "About three blocks, down that way."

You look in the direction he points and see nothing familiar. Looking the other way, you see a policeman about a block away writing a parking ticket. What do you do?

If you're like most people, you hurry to the policeman to pose the same question. When the policeman says, "About four blocks back the way you came," you realize that the child had been correct, despite the fact that the information coming from him was not sufficiently believable to act upon.



One probably wouldn't feel comfortable accepting the diagnosis of a rare disease made because the doctor couldn't find any other explanation for the symptoms.

The acceptance of an hypothesis as being true is not based just on its resistance to refutation. It is based on the hypothesis's believability, a human requirement usually determined by the prevailing paradigm.

In the 19th century, most scientists thought combustible materials contained "phlogiston," which was released quickly during combustion and slowly during corrosion. Experiments were designed to measure the amount of phlogiston released from iron during rusting by weighing the iron before and after the rusting process. The experimenters found that the rusted iron weighed slightly more, not slightly less, than before. The weight increase resulted from chemical combination of oxygen with the iron. The finding was assumed to be an error of measurement, even though it was repeatable. In the paradigm of the time, the truth was not believable.

What is a silly position in one generation may be a commonly held belief in another. Sir Austin Bradford Hill refers to a 19th century prize-winning essayist, writing on the value and fallacy of statistics, who listed some "absurd" associations and scoffed that "it would be no more ridiculous for the stranger who passed the night in the steerage of an emigrant ship to ascribe the typhus, which he there contracted, to the vermin with which bodies of the sick might be infected."

When is a Fact a Fact and Hormone Replacement Therapy

An important human factor influencing the application of the scientific method is deciding how many times to further test an hypothesis that has not yet been refuted. The more times and the more different ways an hypothesis is tested, the more likely an untrue hypothesis will be weeded out. Medical researchers in the mid-20th century noted that women were at relatively low risk for death from heart disease, compared to males, until after menopause. Researchers interpreted that these lower rates signified that premenopausal levels of the female hormone, estrogen, protected women from coronary heart disease. Several studies appeared to support that conclusion. Clinicians, pushed by an aggressive marketing program by pharmaceutical manufacturers, regarded the hypothesis as fact when making the decision for recommending estrogen replacement therapy (ERT) for women going through menopause and for the years after. When it was found in 1975 that ERT alone (without progesterone) increased the risk of uterine cancer, progesterone was added to the recommended formula, which remedied the unacceptable risk of endometrial cancer.

Because of a need to evaluate all the benefits and risks, including the risk of cancer of the breast, another estrogen-dependent organ, the National Heart, Lung, and Blood Institute funded 40 collaborating clinical institutions in the U.S. to undertake clinical trials to assess the risks and benefits of ERT. Recently, researchers halted the trial in the U.S. because the study revealed that women receiving the ERT developed significantly higher rates of heart disease than those receiving the placebo treatment.

At that time, it was not known that typhus was transmitted by the bite of an infected louse.

In the mid-20th century Yerushalmy recorded the birth weights of all babies born to mothers in a large HMO, and whether or not their mothers or fathers smoked cigarettes during the pregnancy. He found that children born of smoking mothers were more likely to be smaller at birth. However, when he also found that newborns of non-smoking mothers were also likely to be smaller if the father smoked, he doubted that there was a biological reason for the low birth weight in both instances. At that time, the effect of exposure to second-hand smoke was not yet recognized. *(Please see Biological plausibility in the Seven Criteria for a Causal Association for another interesting example. -ed.)*

Bias

One important way in which an untrue hypothesis can appear to be supported by a study is when an association does not really exist, but an observed association gains credence because of an error on the part of the investigator. We call this type of error "bias".

Bias is non-random error. It can create the illusion of an association when, in fact, none exists. It results from errors in the design of a study, in the measurements gathered, or in the analysis conducted. If the researcher repeats an error of this or a similar kind, he or she will continue to get similar results. There are several strategies for eliminating bias as an explanation. One is to identify potential sources and design ways to avoid them. For example, if one is concerned that an investigative team may inadvertently interpret measurements from cases differently from measurements of controls, "blinding" the investigator who interprets the study's data can avoid that particular bias.

A famous example of bias involves a very prestigious team of researchers investigating risk factors for pancreatic cancer. The team found that members of the cancer case group had a history of drinking coffee that far exceeded the amount of coffee drunk by the control group (those known not to have the disease) selected for the study. The cases (those known to have the disease) were recruited as patients from several Massachusetts hospitals, and the control participants were patients of the same age and sex recruited from the same hospitals at the same time as the cases.

The results made national headlines and coffee sales slumped. Separate studies using cases recruited from among previously diagnosed patients in a particular geographic area, and controls randomly selected from the same population, failed to find that coffee had been drunk more by cases than by controls. It was later determined that in the hospital-based study, many patient controls had conditions such as peptic ulcers in which drinking coffee made the symptoms worse. These controls therefore tended not to drink coffee. An unwitting bias in the selection of the hospital controls made it appear that an association existed between coffee drinking and pancreatic cancer.

Reverse Causality

All possible explanations for an observed association can be affected by human judgment. In "reverse causality", an association really does exist, but the cause and effect are reversed. This is usually not a difficult alternative to evaluate. For one thing, the

cause has to happen before the effect. However, if the suspected cause and effect are measured simultaneously, such as blood cholesterol level and narrowed coronary arteries, one might not be able to tell which came first.

Confounding

Still another explanation for an observed association is termed "confounding." A confounder is a real causal factor that happens to be associated with the suspected causal factor, and both have an association with the outcome of interest. This is true "guilt by association." An example is the finding that fathers of children born with Down syndrome are older than fathers of other newborns. It is not because of any biological effect of being an older father, however. It is because older fathers tend to be married to older mothers and older mothers do have a biological reason for being at higher risk of giving birth to a child with Down syndrome. If one studies fathers as the factor of interest, one will find an association between the age of the father and the risk of Down syndrome. The age of the mother, however, causes the effect and happens to be associated with the suspected factor (father's age).

The Test of Truth

Sherlock Holmes advised Dr. Watson, "when you have eliminated the impossible, whatever remains, however improbable, must be the truth." In epidemiological research, one can't always determine that alternative explanations are impossible, but often they can be determined to be extremely implausible and unlikely. If it is possible to reasonably discount the possibilities of chance, bias, reverse causality,

Seven Criteria for a Causal Association

1. **Strength.** Generally, the stronger the association, the more likely it is to be causal. The smaller the association, the more likely it is to be caused by some unrecognized error. There are two ways of judging an association's strength. One is by how large the association is (e.g., a large relative risk or correlation coefficient). For example, the relative risk of a rare vaginal cancer among daughters of women who were prescribed DES during their early pregnancy is about 400 times that of daughters not so exposed in utero. It is unlikely that an unrecognized error could produce so strong a finding. The other is how likely the association is to have occurred by chance (e.g., 1 chance in a million). Ideally, both interpretations of strength are present.

2. **Consistency.** There are two components to consistency, internal and external. The first is the similarity of the finding in various subgroups in a study, or closely related associations in the same study (internal consistency). In a study of oral contraceptives (OCs) and the risk of breast cancer, Paffenbarger found no association when looking at all women as a group. When individual subgroups were examined, however, one subgroup was found to be at higher risk: women who had never been pregnant, and who had taken OCs longer than 2 years but fewer 7 years. No other groups had elevated risks. If one accepts the supposition that the observed association is causal, then one has to explain why taking OCs longer than 7 years reverses the causal effect. One would also need to explain why OCs cause breast cancer only among women who had never been pregnant, but not in others. This example thus has an internal inconsistency and the association probably resulted from chance. If all subgroups had an elevated risk, it would evidence internal consistency.

In a study of an epidemic of unexpected hospital deaths due to cardiac arrest, a strong association was found with a particular nurse being on duty. In addition, other associations with the unexpected cardiac arrests were found with the following (all consistent with but different from the first): night shift after the ward lights were turned off, being a patient in a particular bed not in sight of the nurses' station, and having an IV running. If the nurse association is not causal, what is the explanation for the others?

The second component (external consistency) is the similarity of study findings with other studies of the same thing. In other words, the findings are repeatable. Repeatable results are not the result of chance. It is particularly convincing when similar results are obtained by different investigators, studying different populations, using different methods. Similar errors in design, analysis, or measurement are unlikely to be present among all the different studies.

3. **Specificity.** Traditionally, the finding of a factor with multiple effects was suspected of being an error. Early investigators reasoned that if the calculated rate of cancer in a group was elevated, but so was the rate of heart disease, stroke, infectious disease, accidents, and suicide, the explanation was more likely to be a miscount of the number of people in the group, rather than all the rates being elevated. However, since we now recognize that use of tobacco can cause various types of cancers, heart disease, stroke, chronic lung disease, and even wrinkles, that interpretation is now largely discounted. More recently,

Continued on next page…

Continued from previous page…

epidemiologists have learned that when a risk factor is narrowly defined, both with respect to the hypothesis being tested and to the definitions of the risk factor(s) and the outcomes, this criterion can be used. An association with the hypothesized cause and effect very specifically defined is unlikely to suffer from a mixture of similar but different factors and effects.

For example, in the early 1970's, researchers lumped together a number of sudden unexplained infant deaths under the rubric Sudden Infant Death Syndrome (SIDS). However, in the mid-70's, some investigators identified a cause for some of the cases. This cause was "infant botulism." Before that time it was not recognized that the spores of the bacterium Clostridium botulinum could germinate and produce the botulism toxin in the gut of an infant. (Adults are not susceptible to that particular event.) Raw honey commonly contains spores of C. botulinum. Parents who added raw unpasteurized honey to their infant's formula unwittingly placed their infant at risk for this type of rapid onset death. After that cause was identified, it was possible to exclude those deaths from the general category of SIDS. In so doing, the SIDS category became a little more homogeneous, making the condition easier to study. (Sleeping position, parental smoking, etc., had no relationship to infant botulism deaths, of course.) At the start of the research several seemingly similar but unrelated effects, each a with different cause, were lumped together. Application of the criterion of specificity helped distinguish cause-and-effect relationships.

4. **Biological plausibility.** When the association apparently operates through known biological mechanisms, the factor is more likely to be causal. Biological plausibility is the most treacherous criterion, because the state of accepted knowledge changes over time and plausibility is dependent upon the state of knowledge. This criterion contributes most to the believability of an association, but usually contributes the least to solid evidence for or against its causality.

Several decades ago, it was accepted that most chronic diseases (e.g., cancer, heart disease, stroke, ulcerative colitis) were multifactorial, and infectious organisms were not among the multiple causal factors. Stomach ulcers were attributed to some combination of smoking, stress, alcohol use, diet and constitutional factors. When a New Zealand graduate student noted that a particular bacterium could be found in the stomach mucosa of some patients with ulcers, he conducted a study of patients with and without ulcers. He detected a large difference in the prevalence of Helicobacter organisms between the two groups. The finding was met with mild interest, great skepticism, and some ridicule of the idea that the presence of a bacterium could be causally related to stomach ulcers. In the paradigm of the day, that concept was not biologically plausible. Today, Helicobacter is accepted as a cause of both stomach ulcers and stomach cancer.

5. **Temporality.** Conceptually, this is nearly the same as reverse causality. If something is a cause, it must occur before, and not after, the effect or outcome.

6. **A dose-response effect.** Increased risk with increased exposure or dose is a typical characteristic of a cause-and-effect relationship, and when it is present, it greatly increases the probability that an observed association is a cause.

7. **Intervention effect.** If the removal of the factor is accompanied by the removal of the outcome, it is convincing evidence of a cause-and-effect relationship.

and confounding as explanations for an observed association, logically only one possible explanation remains: the cause. Notwithstanding Sherlock Holmes, however, such a conclusion is based only on negative evidence. Scientists must look for positive evidence to ensure that the association is truly causal. One probably wouldn't feel comfortable accepting the diagnosis of a rare disease made because the doctor couldn't find any other explanation for the symptoms, but would be more accepting if informed that some reliable positive test results indicated the presence of the disease.

There are seven tests (criteria) that serve to build and strengthen a conclusion of causality if present, or weaken and demolish a causal conclusion if absent. Typically, in a true cause-effect relationship, most of the seven are clearly present. Their discovery depends upon their being sought; sometimes there is simply a lack of evidence, pro or con. These are described in "Seven Criteria for a Casual Association (sidebar)."

In epidemiology, sometimes a study only has access to data that don't allow consideration of all seven of the criteria. The data available to the study may have been collected at the same time, so that a temporal relationship between the suspected cause and the labeled effect isn't available. Sometimes the number of available subjects for a study is not large enough to look for increments of exposure; the dose-response effect can't be evaluated. For example, in a study of smoking and the risk of lung cancer with only 60 cases, one could probably lump all smokers together and compare them to

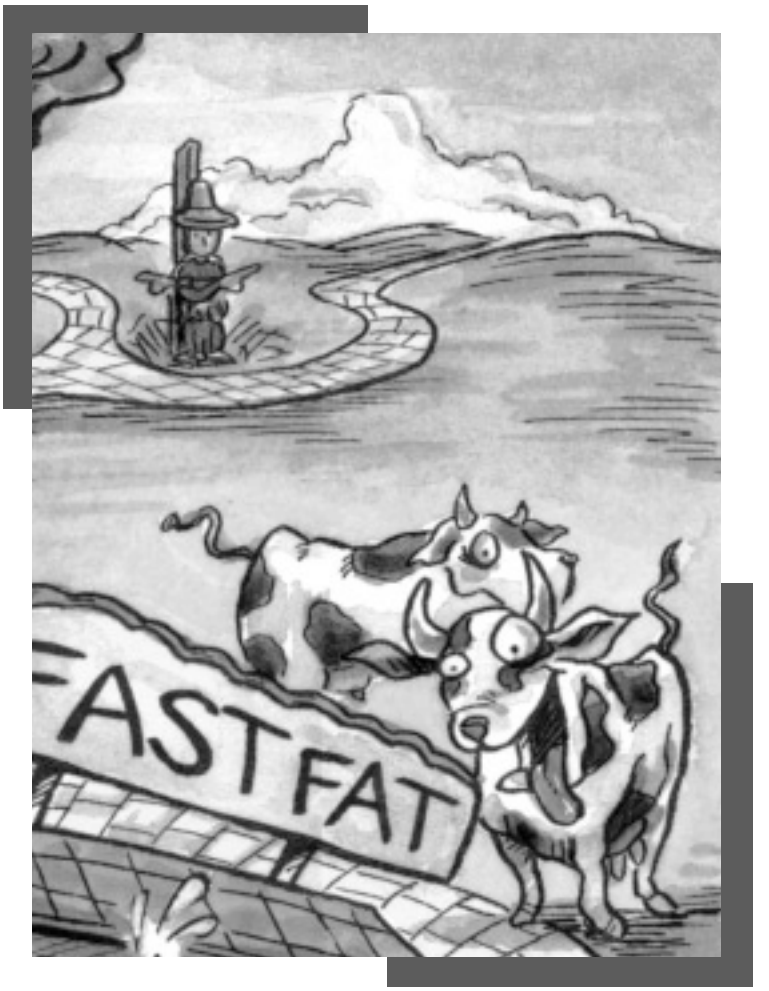
> **"It is easy to arrive at the right decision given enough time and enough information. But there is seldom enough information and there is never enough time."**
>
> **John F. Kennedy**

non-smokers. It may not be possible to compare two-pack-a-day smokers to one-pack-a-day smokers to see if they had a higher risk. Certainly, one would not be able to compare those who smoke 1, 2, 3, 4, 5, 6, 7, etc., cigarettes a day because there simply would not be enough cases in each category. Such a situation requires one to evaluate based on the data available, i.e., without complete information. This is a typical situation in epidemiological studies.

The Real World and Public Policy

Sometimes urgent situations demand that epidemiologists and policy makers make decisions based not solely on the criteria for a causal relationship, but also on human judgment and sometimes human prejudices, as well as values that place a premium on safety. This leads to situations in which governing agencies and individuals must make public health decisions based on incomplete or imperfect information. As John F. Kennedy remarked, "It is easy to arrive at the right decision given enough time and enough information. But there is seldom enough information and there is never enough time." The following example illustrates many of the issues that epidemiologist and policy makers commonly face.

In 1989, an amino acid, L-tryptophan (LT) had become quite popular as a natural tranquilizer and sleep inducer. It was considered safe because it is one of the amino acids essential to complete nutrition for humans, and is part of our normal diet. (The soporific effect of a warm glass of milk before bedtime is attributed to milk's high content of this amino acid.) Drug and health food stores sold LT over the counter under a number of different labels. Psychiatrists and other physicians recommended it to patients with anxiety or insomnia. Then, quite suddenly, there were outbreaks of a serious new medical condition, sometimes leading to death, characterized by extreme muscle inflammation and soreness, disabling tiredness, and counts of eosinophils (a white blood cell usually associated with allergies) at levels usually seen only in cases of the rare eosinophilic leukemia. Health officials in several states, including Oregon, rapidly conducted



studies on an urgent basis. They all found that those with this new condition (eosinophilic myalgia syndrome, or EMS) were almost exclusively limited to those who took LT, whereas LT use was uncommon in selected controls. With further investigation of LT users only, investigators found the EMS to be strongly associated with LT from a single manufacturer, Showa Denko, which had been sold under nearly a dozen labels. Further, investigators found the association only with LT produced after that manufacturer changed its production process. As a result of these findings, the government took all sources of LT off the market.

This episode demonstrates all the decisions and evaluations necessary in considering epidemiologic findings. First, the findings were repeatable in several studies, though there was some argument that at least in some studies, the selection of controls was biased. Second, the association with one manufacturer's LT (produced after a process change) supported the first association. Third, although it could not be absolutely established that most EMS patients took LT before developing EMS, the alternative explanation (that people with EMS sought relief from early symptoms by purchasing LT from a single manufacturer) was untenable. Fourth, the size of the risk of EMS and the specificity of the LT to a single manufacturer strengthened the hypothesis that something had contaminated the new manufacturing process.

However, investigators could not offer "absolute proof" of the causal nature of the association. Conducting a clinical experiment with some people receiving LT from the suspected manufacturer and others receiving a placebo would be impossibly unethical. Government officials made the decision on the available evidence; more research was not an acceptable option.

The Food and Drug Administration decided that evidence was sufficiently strong to consider the association causal, and that any harm done by action on their part would be far less than the possible harm if they failed to act. The association fulfilled most of the epidemiologic criteria for a cause, even though all of the seven criteria for supporting a cause (notably, an intervention effect and an indisputable dose-response effect) were simply not available. The outbreak of EMS ended with the FDA action.

In this case the FDA decision withstood the criticisms of those whose "ox was gored".

This issue remains controversial. Physicians in the US can still legally prescribe L-tryptophan and many people in the natural supplement industry believed that the FDA banned L-tryptophan from over-the-counter sales for political reasons due to influence by the pharmaceutical industry. However, this real world example of a public health intervention is not an example of a frozen scientific paradigm as some would believe.

The subsequent re-introduction of LT as a prescription item can be debated as to its appropriateness. Would the banning of over-the-counter LT in the US have been as likely were there not other (prescription) remedies for the same problems? Probably not, but it was apparent that sufferers of depression, anxiety, and insomnia already had other remedies that had passed the FDA criteria for being safe and effective. Were drug companies happy about that? Undoubtedly so. Was there some clandestine government-industry agreement to support the development of new drugs? Possibly, but one doesn't have to posit a conspiracy to rationalize the action.

Most people reviewing the evidence would support the FDA action. I pose this question each year to graduate classes of about 18-20 students (most of whom already have one doctoral degree) who must review the original articles with respect to causation and to a justified resultant policy. Most agree with the total ban policy, but each year several (10-15%) think LT should be generally available over the counter, except not if it is manufactured using the same process as the Showa Denko lots. The only question considered in the investigation was whether the evidence about LT met the criteria for causation, given the evidence at the time and the consequences of making an error. The primary agenda in this case was public safety and public safety was served by the FDA's decision.

Don Austin is Professor, Department of Public Health and Preventive Medicine, at OHSU. He is a medical epidemiologist with primary interests in cancer and the quality of care. He served in state and federal health agencies for several decades before coming to academia. As a teacher, his favorite topic is that of determining cause.