http://www.nytimes.com/2007/09/16/magazine/16epidemiology-t.html?_r=1&pagewanted=all&oref=slogin

Do We Really Know What Makes Us Healthy?

Reinhard Hunger

By GARY TAUBES
Published: September 16, 2007

Once upon a time, women took estrogen only to relieve the hot flashes, sweating, vaginal dryness and the other discomforting symptoms of menopause. In the late 1960s, thanks in part to the efforts of Robert Wilson, a Brooklyn gynecologist, and his 1966 best seller, “Feminine Forever,” this began to change, and estrogen therapy evolved into a long-term remedy for the chronic ills of aging. Menopause, Wilson argued, was not a natural age-related condition; it was an illness, akin to diabetes or kidney failure, and one that could be treated by taking estrogen to replace the hormones that a woman’s ovaries secreted in ever diminishing amounts. With this argument estrogen evolved into hormone-replacement therapy, or H.R.T., as it came to be called, and became one of the most popular prescription drug treatments in America.

By the mid-1990s, the American Heart Association, the American College of Physicians and the American College of Obstetricians and Gynecologists had all concluded that the beneficial effects of H.R.T. were sufficiently well established that it could be recommended to older women as a means of warding off heart disease and osteoporosis. By 2001, 15 million women were filling H.R.T. prescriptions annually; perhaps 5 million were older women, taking the drug solely with the expectation that it

would allow them to lead a longer and healthier life. A year later, the tide would turn. In the summer of 2002, estrogen therapy was exposed as a hazard to health rather than a benefit, and its story became what Jerry Avorn, a Harvard epidemiologist, has called the “estrogen debacle” and a “case study waiting to be written” on the elusive search for truth in medicine.

Many explanations have been offered to make sense of the here-today-gone-tomorrow nature of medical wisdom — what we are advised with confidence one year is reversed the next — but the simplest one is that it is the natural rhythm of science. An observation leads to a hypothesis. The hypothesis (last year’s advice) is tested, and it fails this year’s test, which is always the most likely outcome in any scientific endeavor. There are, after all, an infinite number of wrong hypotheses for every right one, and so the odds are always against any particular hypothesis being true, no matter how obvious or vitally important it might seem.

In the case of H.R.T., as with most issues of diet, lifestyle and disease, the hypotheses begin their transformation into public-health recommendations only after they’ve received the requisite support from a field of research known as epidemiology. This science evolved over the last 250 years to make sense of epidemics — hence the name — and infectious diseases. Since the 1950s, it has been used to identify, or at least to try to identify, the causes of the common chronic diseases that befall us, particularly heart disease and cancer. In the process, the perception of what epidemiologic research can legitimately accomplish — by the public, the press and perhaps by many epidemiologists themselves — may have run far ahead of the reality. The case of hormone-replacement therapy for post-menopausal women is just one of the cautionary tales in the annals of epidemiology. It’s a particularly glaring example of the difficulties of trying to establish reliable knowledge in any scientific field with research tools that themselves may be unreliable.

What was considered true about estrogen therapy in the 1960s and is still the case today is that it is an effective treatment for menopausal symptoms. Take H.R.T. for a few menopausal years and it’s extremely unlikely that any harm will come from it. The uncertainty involves the lifelong risks and benefits should a woman choose to continue taking H.R.T. long past menopause. In 1985, the Nurses’ Health Study run out of the Harvard Medical School and the Harvard School of Public Health reported that women taking estrogen had only a third as many heart attacks as women who had never taken the drug. This appeared to confirm the belief that women were protected from heart attacks until they passed through menopause and that it was estrogen that bestowed that protection, and this became the basis of the therapeutic wisdom for the next 17 years.

Faith in the protective powers of estrogen began to erode in 1998, when a clinical trial called HERS, for Heart and Estrogen-progestin Replacement Study, concluded that estrogen therapy increased, rather than decreased, the likelihood that women who already had heart disease would suffer a heart attack. It evaporated entirely in July 2002, when a second trial, the Women’s Health Initiative, or W.H.I., concluded that

H.R.T. constituted a potential health risk for all postmenopausal women. While it might protect them against osteoporosis and perhaps colorectal cancer, these benefits would be outweighed by increased risks of heart disease, stroke, blood clots, breast cancer and perhaps even dementia. And that was the final word. Or at least it was until the June 21 issue of The New England Journal of Medicine. Now the idea is that hormone-replacement therapy may indeed protect women against heart disease if they begin taking it during menopause, but it is still decidedly deleterious for those women who begin later in life.

This latest variation does come with a caveat, however, which could have been made at any point in this history. While it is easy to find authority figures in medicine and public health who will argue that today's version of H.R.T. wisdom is assuredly the correct one, it's equally easy to find authorities who will say that surely we don't know. The one thing on which they will all agree is that the kind of experimental trial necessary to determine the truth would be excessively expensive and time-consuming and so will almost assuredly never happen. Meanwhile, the question of how many women may have died prematurely or suffered strokes or breast cancer because they were taking a pill that their physicians had prescribed to protect them against heart disease lingers unanswered. A reasonable estimate would be tens of thousands.

The Flip-Flop Rhythm of Science

At the center of the H.R.T. story is the science of epidemiology itself and, in particular, a kind of study known as a prospective or cohort study, of which the Nurses' Health Study is among the most renowned. In these studies, the investigators monitor disease rates and lifestyle factors (diet, physical activity, prescription drug use, exposure to pollutants, etc.) in or between large populations (the 122,000 nurses of the Nurses' study, for example). They then try to infer conclusions — i.e., hypotheses — about what caused the disease variations observed. Because these studies can generate an enormous number of speculations about the causes or prevention of chronic diseases, they provide the fodder for much of the health news that appears in the media — from the potential benefits of fish oil, fruits and vegetables to the supposed dangers of sedentary lives, trans fats and electromagnetic fields. Because these studies often provide the only available evidence outside the laboratory on critical issues of our well-being, they have come to play a significant role in generating public-health recommendations as well.

The dangerous game being played here, as David Sackett, a retired Oxford University epidemiologist, has observed, is in the presumption of preventive medicine. The goal of the endeavor is to tell those of us who are otherwise in fine health how to remain healthy longer. But this advice comes with the expectation that any prescription given — whether diet or drug or a change in lifestyle — will indeed prevent disease rather than be the agent of our disability or untimely death. With that presumption, how unambiguous does the evidence have to be before any advice is offered?

The catch with observational studies like the Nurses' Health Study, no matter how well designed and how many tens of thousands of subjects they might include, is that they have a fundamental limitation. They can distinguish associations between two events — that women who take H.R.T. have less heart disease, for instance, than women who don't. But they cannot inherently determine causation — the conclusion that one event causes the other; that H.R.T. protects against heart disease. As a result, observational studies can only provide what researchers call hypothesis-generating evidence — what a defense attorney would call circumstantial evidence.

Testing these hypotheses in any definitive way requires a randomized-controlled trial — an experiment, not an observational study — and these clinical trials typically provide the flop to the flip-flop rhythm of medical wisdom. Until August 1998, the faith that H.R.T. prevented heart disease was based primarily on observational evidence, from the Nurses' Health Study most prominently. Since then, the conventional wisdom has been based on clinical trials — first HERS, which tested H.R.T. against a placebo in 2,700 women with heart disease, and then the Women's Health Initiative, which tested the therapy against a placebo in 16,500 healthy women. When the Women's Health Initiative concluded in 2002 that H.R.T. caused far more harm than good, the lesson to be learned, wrote Sackett in The Canadian Medical Association Journal, was about the "disastrous inadequacy of lesser evidence" for shaping medical and public-health policy. The contentious wisdom circa mid-2007 — that estrogen benefits women who begin taking it around the time of menopause but not women who begin substantially later — is an attempt to reconcile the discordance between the observational studies and the experimental ones. And it may be right. It may not. The only way to tell for sure would be to do yet another randomized trial, one that now focused exclusively on women given H.R.T. when they begin their menopause.

A Poor Track Record of Prevention

No one questions the value of these epidemiologic studies when they're used to identify the unexpected side effects of prescription drugs or to study the progression of diseases or their distribution between and within populations. One reason researchers believe that heart disease and many cancers can be prevented is because of observational evidence that the incidence of these diseases differ greatly in different populations and in the same populations over time. Breast cancer is not the scourge among Japanese women that it is among American women, but it takes only two generations in the United States before Japanese-Americans have the same breast cancer rates as any other ethnic group. This tells us that something about the American lifestyle or diet is a cause of breast cancer. Over the last 20 years, some two dozen large studies, the Nurses' Health Study included, have so far failed to identify what that factor is. They may be inherently incapable of doing so. Nonetheless, we know that such a carcinogenic factor of diet or lifestyle exists, waiting to be identified.

These studies have also been invaluable for identifying predictors of disease — risk factors — and this information can then guide physicians in weighing the risks and benefits of putting a particular patient on a particular drug. The studies have repeatedly

confirmed that high blood pressure is associated with an increased risk of heart disease and that obesity is associated with an increased risk of most of our common chronic diseases, but they have not told us what it is that raises blood pressure or causes obesity. Indeed, if you ask the more skeptical epidemiologists in the field what diet and lifestyle factors have been convincingly established as causes of common chronic diseases based on observational studies without clinical trials, you'll get a very short list: smoking as a cause of lung cancer and cardiovascular disease, sun exposure for skin cancer, sexual activity to spread the papilloma virus that causes cervical cancer and perhaps alcohol for a few different cancers as well.

Richard Peto, professor of medical statistics and epidemiology at Oxford University, phrases the nature of the conflict this way: "Epidemiology is so beautiful and provides such an important perspective on human life and death, but an incredible amount of rubbish is published," by which he means the results of observational studies that appear daily in the news media and often become the basis of public-health recommendations about what we should or should not do to promote our continued good health.

In January 2001, the British epidemiologists George Davey Smith and Shah Ebrahim, co-editors of The International Journal of Epidemiology, discussed this issue in an editorial titled "Epidemiology — Is It Time to Call It a Day?" They noted that those few times that a randomized trial had been financed to test a hypothesis supported by results from these large observational studies, the hypothesis either failed the test or, at the very least, the test failed to confirm the hypothesis: antioxidants like vitamins E and C and beta carotene did not prevent heart disease, nor did eating copious fiber protect against colon cancer.

The Nurses' Health Study is the most influential of these cohort studies, and in the six years since the Davey Smith and Ebrahim editorial, a series of new trials have chipped away at its credibility. The Women's Health Initiative hormone-therapy trial failed to confirm the proposition that H.R.T. prevented heart disease; a W.H.I. diet trial with 49,000 women failed to confirm the notion that fruits and vegetables protected against heart disease; a 40,000-woman trial failed to confirm that a daily regimen of low-dose aspirin prevented colorectal cancer and heart attacks in women under 65. And this June, yet another clinical trial — this one of 1,000 men and women with a high risk of colon cancer — contradicted the inference from the Nurses's study that folic acid supplements reduced the risk of colon cancer. Rather, if anything, they appear to increase risk.

The implication of this track record seems hard to avoid. "Even the Nurses' Health Study, one of the biggest and best of these studies, cannot be used to reliably test small-to-moderate risks or benefits," says Charles Hennekens, a principal investigator with the Nurses' study from 1976 to 2001. "None of them can."

Proponents of the value of these studies for telling us how to prevent common diseases — including the epidemiologists who do them, and physicians, nutritionists and public-

health authorities who use their findings to argue for or against the health benefits of a particular regimen — will argue that they are never relying on any single study. Instead, they base their ultimate judgments on the "totality of the data," which in theory includes all the observational evidence, any existing clinical trials and any laboratory work that might provide a biological mechanism to explain the observations.

This in turn leads to the argument that the fault is with the press, not the epidemiology. "The problem is not in the research but in the way it is interpreted for the public," as Jerome Kassirer and Marcia Angell, then the editors of The New England Journal of Medicine, explained in a 1994 editorial titled "What Should the Public Believe?" Each study, they explained, is just a "piece of a puzzle" and so the media had to do a better job of communicating the many limitations of any single study and the caveats involved — the foremost, of course, being that "an association between two events is not the same as a cause and effect."

Stephen Pauker, a professor of medicine at Tufts University and a pioneer in the field of clinical decision making, says, "Epidemiologic studies, like diagnostic tests, are probabilistic statements." They don't tell us what the truth is, he says, but they allow both physicians and patients to "estimate the truth" so they can make informed decisions. The question the skeptics will ask, however, is how can anyone judge the value of these studies without taking into account their track record? And if they take into account the track record, suggests Sander Greenland, an epidemiologist at the University of California, Los Angeles, and an author of the textbook "Modern Epidemiology," then wouldn't they do just as well if they simply tossed a coin?

As John Bailar, an epidemiologist who is now at the National Academy of Science, once memorably phrased it, "The appropriate question is not whether there are uncertainties about epidemiologic data, rather, it is whether the uncertainties are so great that one cannot draw useful conclusions from the data."

Science vs. the Public Health

Understanding how we got into this situation is the simple part of the story. The randomized-controlled trials needed to ascertain reliable knowledge about long-term risks and benefits of a drug, lifestyle factor or aspect of our diet are inordinately expensive and time consuming. By randomly assigning research subjects into an intervention group (who take a particular pill or eat a particular diet) or a placebo group, these trials "control" for all other possible variables, both known and unknown, that might effect the outcome: the relative health or wealth of the subjects, for instance. This is why randomized trials, particularly those known as placebo-controlled, double-blind trials, are typically considered the gold standard for establishing reliable knowledge about whether a drug, surgical intervention or diet is really safe and effective.

But clinical trials also have limitations beyond their exorbitant costs and the years or decades it takes them to provide meaningful results. They can rarely be used, for instance, to study suspected harmful effects. Randomly subjecting thousands of

individuals to secondhand tobacco smoke, pollutants or potentially noxious trans fats presents obvious ethical dilemmas. And even when these trials are done to study the benefits of a particular intervention, it's rarely clear how the results apply to the public at large or to any specific patient. Clinical trials invariably enroll subjects who are relatively healthy, who are motivated to volunteer and will show up regularly for treatments and checkups. As a result, randomized trials "are very good for showing that a drug does what the pharmaceutical company says it does," David Atkins, a preventive-medicine specialist at the Agency for Healthcare Research and Quality, says, "but not very good for telling you how big the benefit really is and what are the harms in typical people. Because they don't enroll typical people."

These limitations mean that the job of establishing the long-term and relatively rare risks of drug therapies has fallen to observational studies, as has the job of determining the risks and benefits of virtually all factors of diet and lifestyle that might be related to chronic diseases. The former has been a fruitful field of research; many side effects of drugs have been discovered by these observational studies. The latter is the primary point of contention.

While the tools of epidemiology — comparisons of populations with and without a disease — have proved effective over the centuries in establishing that a disease like cholera is caused by contaminated water, as the British physician John Snow demonstrated in the 1850s, it's a much more complicated endeavor when those same tools are employed to elucidate the more subtle causes of chronic disease.

And even the success stories taught in epidemiology classes to demonstrate the historical richness and potential of the field — that pellagra, a disease that can lead to dementia and death, is caused by a nutrient-deficient diet, for instance, as Joseph Goldberger demonstrated in the 1910s — are only known to be successes because the initial hypotheses were subjected to rigorous tests and happened to survive them. Goldberger tested the competing hypothesis, which posited that the disease was caused by an infectious agent, by holding what he called "filth parties," injecting himself and seven volunteers, his wife among them, with the blood of pellagra victims. They remained healthy, thus doing a compelling, if somewhat revolting, job of refuting the alternative hypothesis.

Smoking and lung cancer is the emblematic success story of chronic-disease epidemiology. But lung cancer was a rare disease before cigarettes became widespread, and the association between smoking and lung cancer was striking: heavy smokers had 2,000 to 3,000 percent the risk of those who had never smoked. This made smoking a "turkey shoot," says Greenland of U.C.L.A., compared with the associations epidemiologists have struggled with ever since, which fall into the tens of a percent range. The good news is that such small associations, even if causal, can be considered relatively meaningless for a single individual. If a 50-year-old woman with a small risk of breast cancer takes H.R.T. and increases her risk by 30 percent, it remains a small risk.

The compelling motivation for identifying these small effects is that their impact on the public health can be enormous if they're aggregated over an entire nation: if tens of millions of women decrease their breast cancer risk by 30 percent, tens of thousands of such cancers will be prevented each year. In fact, between 2002 and 2004, breast cancer incidence in the United States dropped by 12 percent, an effect that may have been caused by the coincident decline in the use of H.R.T. (And it may not have been. The coincident reduction in breast cancer incidence and H.R.T. use is only an association.)

Saving tens of thousands of lives each year constitutes a powerful reason to lower the standard of evidence needed to suggest a cause-and-effect relationship — to take a leap of faith. This is the crux of the issue. From a scientific perspective, epidemiologic studies may be incapable of distinguishing a small effect from no effect at all, and so caution dictates that the scientist refrain from making any claims in that situation. From the public-health perspective, a small effect can be a very dangerous or beneficial thing, at least when aggregated over an entire nation, and so caution dictates that action be taken, even if that small effect might not be real. Hence the public-health logic that it's better to err on the side of prudence even if it means persuading us all to engage in an activity, eat a food or take a pill that does nothing for us and ignoring, for the moment, the possibility that such an action could have unforeseen harmful consequences. As Greenland says, "The combination of data, statistical methodology and motivation seems a potent anesthetic for skepticism."

The Bias of Healthy Users

The Nurses' Health Study was founded at Harvard in 1976 by Frank Speizer, an epidemiologist who wanted to study the long-term effects of oral contraceptive use. It was expanded to include postmenopausal estrogen therapy because both treatments involved long-term hormone use by millions of women, and nobody knew the consequences. Speizer's assistants in this endeavor, who would go on to become the most influential epidemiologists in the country, were young physicians — Charles Hennekens, Walter Willett, Meir Stampfer and Graham Colditz — all interested in the laudable goal of preventing disease more than curing it after the fact.

When the Nurses' Health Study first published its observations on estrogen and heart disease in 1985, it showed that women taking estrogen therapy had only a third the risk of having a heart attack as had women who had never taken it; the association seemed compelling evidence for a cause and effect. Only 90 heart attacks had been reported among the 32,000 postmenopausal nurses in the study, and Stampfer, who had done the bulk of the analysis, and his colleagues "considered the possibility that the apparent protective effect of estrogen could be attributed to some other factor associated with its use." They decided, though, as they have ever since, that this was unlikely. The paper's ultimate conclusion was that "further work is needed to define the optimal type, dose and duration of postmenopausal hormone use" for maximizing the protective benefit.

Only after Stampfer and his colleagues published their initial report on estrogen therapy did other investigators begin to understand the nature of the other factors that might explain the association. In 1987, Diana Petitti, an epidemiologist now at the University of Southern California, reported that she, too, had detected a reduced risk of heart-disease deaths among women taking H.R.T. in the Walnut Creek Study, a population of 16,500 women. When Petitti looked at all the data, however, she "found an even more dramatic reduction in death from homicide, suicide and accidents." With little reason to believe that estrogen would ward off homicides or accidents, Petitti concluded that something else appeared to be "confounding" the association she had observed. "The same thing causing this obvious spurious association might also be contributing to the lower risk of coronary heart disease," Petitti says today.

That mysterious something is encapsulated in what epidemiologists call the healthy-user bias, and some of the most fascinating research in observational epidemiology is now aimed at understanding this phenomenon in all its insidious subtlety. Only then can epidemiologists learn how to filter out the effect of this healthy-user bias from what might otherwise appear in their studies to be real causal relationships. One complication is that it encompasses a host of different and complex issues, many or most of which might be impossible to quantify. As Jerry Avorn of Harvard puts it, the effect of healthy-user bias has the potential for "big mischief" throughout these large epidemiologic studies.

At its simplest, the problem is that people who faithfully engage in activities that are good for them — taking a drug as prescribed, for instance, or eating what they believe is a healthy diet — are fundamentally different from those who don't. One thing epidemiologists have established with certainty, for example, is that women who take H.R.T. differ from those who don't in many ways, virtually all of which associate with lower heart-disease risk: they're thinner; they have fewer risk factors for heart disease to begin with; they tend to be more educated and wealthier; to exercise more; and to be generally more health conscious.

Considering all these factors, is it possible to isolate one factor — hormone-replacement therapy — as the legitimate cause of the small association observed or even part of it? In one large population studied by Elizabeth Barrett-Connor, an epidemiologist at the University of California, San Diego, having gone to college was associated with a 50 percent lower risk of heart disease. So if women who take H.R.T. tend to be more educated than women who don't, this confounds the association between hormone therapy and heart disease. It can give the appearance of cause and effect where none exists.

Another thing that epidemiologic studies have established convincingly is that wealth associates with less heart disease and better health, at least in developed countries. The studies have been unable to establish why this is so, but this, too, is part of the healthy-user problem and a possible confounder of the hormone-therapy story and many of the other associations these epidemiologists try to study. George Davey Smith, who began his career studying how socioeconomic status associates with health, says

one thing this research teaches is that misfortunes “cluster” together. Poverty is a misfortune, and the poor are less educated than the wealthy; they smoke more and weigh more; they’re more likely to have hypertension and other heart-disease risk factors, to eat what’s affordable rather than what the experts tell them is healthful, to have poor medical care and to live in environments with more pollutants, noise and stress. Ideally, epidemiologists will carefully measure the wealth and education of their subjects and then use statistical methods to adjust for the effect of these influences — multiple regression analysis, for instance, as one such method is called — but, as Avorn says, it “doesn’t always work as well as we’d like it to.”

The Nurses’ investigators have argued that differences in socioeconomic status cannot explain the associations they observe with H.R.T. because all their subjects are registered nurses and so this “controls” for variations in wealth and education. The skeptics respond that even if all registered nurses had identical educations and income, which isn’t necessarily the case, then their socioeconomic status will be determined by whether they’re married, how many children they have and their husbands’ income. “All you have to do is look at nurses,” Petitti says. “Some are married to C.E.O.’s of corporations and some are not married and still living with their parents. It cannot be true that there is no socioeconomic distribution among nurses.” Stampfer says that since the Women’s Health Initiative results came out in 2002, the Nurses’ Health Study investigators went back into their data to examine socioeconomic status “to the extent that we could” — looking at measures that might indirectly reflect wealth and social class. “It doesn’t seem plausible” that socioeconomic status can explain the association they observed, he says. But the Nurses’ investigators never published that analysis, and so the skeptics have remained unconvinced.

The Bias of Compliance

A still more subtle component of healthy-user bias has to be confronted. This is the compliance or adherer effect. Quite simply, people who comply with their doctors’ orders when given a prescription are different and healthier than people who don’t. This difference may be ultimately unquantifiable. The compliance effect is another plausible explanation for many of the beneficial associations that epidemiologists commonly report, which means this alone is a reason to wonder if much of what we hear about what constitutes a healthful diet and lifestyle is misconceived.

The lesson comes from an ambitious clinical trial called the Coronary Drug Project that set out in the 1970s to test whether any of five different drugs might prevent heart attacks. The subjects were some 8,500 middle-aged men with established heart problems. Two-thirds of them were randomly assigned to take one of the five drugs and the other third a placebo. Because one of the drugs, clofibrate, lowered cholesterol levels, the researchers had high hopes that it would ward off heart disease. But when the results were tabulated after five years, clofibrate showed no beneficial effect. The researchers then considered the possibility that clofibrate appeared to fail only because the subjects failed to faithfully take their prescriptions.

As it turned out, those men who said they took more than 80 percent of the pills prescribed fared substantially better than those who didn't. Only 15 percent of these faithful "adherers" died, compared with almost 25 percent of what the project researchers called "poor adherers." This might have been taken as reason to believe that clofibrate actually did cut heart-disease deaths almost by half, but then the researchers looked at those men who faithfully took their placebos. And those men, too, seemed to benefit from adhering closely to their prescription: only 15 percent of them died compared with 28 percent who were less conscientious. "So faithfully taking the placebo cuts the death rate by a factor of two," says David Freedman, a professor of statistics at the University of California, Berkeley. "How can this be? Well, people who take their placebo regularly are just different than the others. The rest is a little speculative. Maybe they take better care of themselves in general. But this compliance effect is quite a big effect."

The moral of the story, says Freedman, is that whenever epidemiologists compare people who faithfully engage in some activity with those who don't — whether taking prescription pills or vitamins or exercising regularly or eating what they consider a healthful diet — the researchers need to account for this compliance effect or they will most likely infer the wrong answer. They'll conclude that this behavior, whatever it is, prevents disease and saves lives, when all they're really doing is comparing two different types of people who are, in effect, incomparable.

This phenomenon is a particularly compelling explanation for why the Nurses' Health Study and other cohort studies saw a benefit of H.R.T. in current users of the drugs, but not necessarily in past users. By distinguishing among women who never used H.R.T., those who used it but then stopped and current users (who were the only ones for which a consistent benefit appeared), these observational studies may have inadvertently focused their attention specifically on, as Jerry Avorn says, the "Girl Scouts in the group, the compliant ongoing users, who are probably doing a lot of other preventive things as well."

How Doctors Confound the Science

Another complication to what may already appear (for good reason) to be a hopelessly confusing story is what might be called the prescriber effect. The reasons a physician will prescribe one medication to one patient and another or none at all to a different patient are complex and subtle. "Doctors go through a lot of different filters when they're thinking about what kind of drug to give to what kind of person," says Avorn, whose group at Harvard has spent much of the last decade studying this effect. "Maybe they give the drug to their sickest patients; maybe they give it to the people for whom nothing else works."

It's this prescriber effect, combined with what Avorn calls the eager-patient effect, that is one likely explanation for why people who take cholesterol-lowering drugs called statins appear to have a greatly reduced risk of dementia and death from all causes compared with people who don't take statins. The medication itself is unlikely to be the primary

cause in either case, says Avorn, because the observed associations are “so much larger than the effects that have been seen in randomized-clinical trials.”

If we think like physicians, Avorn explains, then we get a plausible explanation: “A physician is not going to take somebody either dying of metastatic cancer or in a persistent vegetative state or with end-stage neurologic disease and say, ‘Let’s get that cholesterol down, Mrs. Jones.’ The consequence of that, multiplied over tens of thousands of physicians, is that many people who end up on statins are a lot healthier than the people to whom these doctors do not give statins. Then add into that the people who come to the doctor and say, ‘My brother-in-law is on this drug,’ or, ‘I saw it in a commercial,’ or, ‘I want to do everything I can to prevent heart disease, can I now have a statin, please?’ Those kinds of patients are very different from the patients who don’t come in. The coup de grÃ¢ce then comes from the patients who consistently take their medications on an ongoing basis, and who are still taking them two or three years later. Those people are special and unusual and, as we know from clinical trials, even if they’re taking a sugar pill they will have better outcomes.”

The trick to successfully understanding what any association might really mean, Avorn adds, is “being clever.” “The whole point of science is self-doubt,” he says, “and asking could there be another explanation for what we’re seeing.”

H.R.T. and the Plausibility Problem

Until the HERS and W.H.I. trials tested and refuted the hypothesis that hormone-replacement therapy protected women against heart disease, Stampfer, Willett and their colleagues argued that these alternative explanations could not account for what they observed. They had gathered so much information about their nurses, they said, that it allowed them to compare nurses who took H.R.T. and engaged in health-conscious behaviors against women who didn’t take H.R.T. and appeared to be equally health-conscious. Because this kind of comparison didn’t substantially change the size of the association observed, it seemed reasonable to conclude that the association reflected the causal effect of H.R.T. After the W.H.I. results were published, says Stampfer, their faith was shaken, but only temporarily. Clinical trials, after all, also have limitations, and so the refutation of what was originally a simple hypothesis — that H.R.T. wards off heart disease — spurred new hypotheses, not quite so simple, to explain it.

At the moment, at least three plausible explanations exist for the discrepancy between the clinical trial results and those of the Nurses’ Health Study and other observational studies. One is that the associations perceived by the epidemiologic studies were due to healthy-user and prescriber effects and not H.R.T. itself. Women who took H.R.T. had less heart disease than women who didn’t, because women who took H.R.T. are different from women who didn’t take H.R.T. And maybe their physicians are also different. In this case, the trials got the right answer; the observational studies got the wrong answer.

A second explanation is that the observational studies got the wrong answer, but only partly. Here, healthy-user and prescriber effects are viewed as minor issues; the question is whether observational studies can accurately determine if women were really taking H.R.T. before their heart attacks. This is a measurement problem, and one conspicuous limitation of all epidemiology is the difficulty of reliably assessing whatever it is the investigators are studying: not only determining whether or not subjects have really taken a medication or consumed the diet that they reported, but whether their subsequent diseases were correctly diagnosed. “The wonder and horror of epidemiology,” Avorn says, “is that it’s not enough to just measure one thing very accurately. To get the right answer, you may have to measure a great many things very accurately.”

The most meaningful associations are those in which all the relevant factors can be ascertained reliably. Smoking and lung cancer, for instance. Lung cancer is an easy diagnosis to make, at least compared with heart disease. And “people sort of know whether they smoke a full pack a day or half or what have you,” says Graham Colditz, who recently left the Nurses’ study and is now at Washington University School of Medicine in St. Louis. “That’s one of the easier measures you can get.” Epidemiologists will also say they believe in the associations between LDL cholesterol, blood pressure and heart disease, because these biological variables are measured directly. The measurements don’t require that the study subjects fill out a questionnaire or accurately recall what their doctors may have told them.

Even the way epidemiologists frame the questions they ask can bias a measurement and produce an association that may be particularly misleading. If researchers believe that physical activity protects against chronic disease and they ask their subjects how much leisure-time physical activity they do each week, those who do more will tend to be wealthier and healthier, and so the result the researchers get will support their preconceptions. If the questionnaire asks how much physical activity a subject’s job entails, the researchers might discover that the poor tend to be more physically active, because their jobs entail more manual labor, and they tend to have more chronic diseases. That would appear to refute the hypothesis.

The simpler the question or the more objective the measurement the more likely it is that an association may stand in the causal pathway, as these researchers put it. This is why the question of whether hormone-replacement therapy effects heart-disease risk, for instance, should be significantly easier to nail down than whether any aspect of diet does. For a measurement “as easy as this,” says Jamie Robins, a Harvard epidemiologist, “where maybe the confounding is not horrible, maybe you can get it right.” It’s simply easier to imagine that women who have taken estrogen therapy will remember and report that correctly — it’s yes or no, after all — than that they will recall and report accurately what they ate and how much of it over the last week or the last year.

But as the H.R.T. experience demonstrates, even the timing of a yes-or-no question can introduce problems. The subjects of the Nurses’ Health Study were asked if they were

taking H.R.T. every two years, which is how often the nurses were mailed new questionnaires about their diets, prescription drug use and whatever other factors the investigators deemed potentially relevant to health. If a nurse fills out her questionnaire a few months before she begins taking H.R.T., as Colditz explains, and she then has a heart attack, say, six months later, the Nurses' study will classify that nurse as "not using" H.R.T. when she had the heart attack.

As it turns out, 40 percent of women who try H.R.T. stay on it for less than a year, and most of the heart attacks recorded in the W.H.I. and HERS trials occurred during the first few years that the women were prescribed the therapy. So it's a reasonable possibility that the Nurses' Health Study and other observational studies misclassified many of the heart attacks that occurred among users of hormone therapy as occurring among nonusers. This is the second plausible explanation for why these epidemiologic studies may have erroneously perceived a beneficial association of hormone use with heart disease and the clinical trials did not.

In the third explanation, the clinical trials and the observational studies both got the right answer, but they asked different questions. Here the relevant facts are that the women who took H.R.T. in the observational studies were mostly younger women going through menopause. Most of the women enrolled in the clinical trials were far beyond menopause. The average age of the women in the W.H.I. trial was 63 and in HERS it was 67. The primary goal of these clinical trials was to test the hypothesis that H.R.T. prevented heart disease. Older women have a higher risk of heart disease, and so by enrolling women in their 60s and 70s, the researchers didn't have to wait nearly as long to see if estrogen protected against heart disease as they would have if they only enrolled women in their 50s.

This means the clinical trials were asking what happens when older women were given H.R.T. years after menopause. The observational studies asked whether H.R.T. prevented heart disease when taken by younger women near the onset of menopause. A different question. The answer, according to Stampfer, Willett and their colleagues, is that estrogen protects those younger women — perhaps because their arteries are still healthy — while it induces heart attacks in the older women whose arteries are not. "It does seem clear now," Willett says, "that the observational studies got it all right. The W.H.I. also got it right for the question they asked: what happens if you start taking hormones many years after menopause? But that is not the question that most women have cared about."

This last explanation is now known as the "timing" hypothesis, and it certainly seems plausible. It has received some support from analyses of small subsets of the women enrolled in the W.H.I. trial, like the study published in June in The New England Journal of Medicine. The dilemma at the moment is that the first two explanations are also plausible. If the compliance effect can explain why anyone faithfully following her doctor's orders will be 50 percent less likely to die over the next few years than someone who's not so inclined, then it's certainly possible that what the Nurses' Health Study and other observational studies did is observe a compliance effect and mistake it

for a beneficial effect of H.R.T. itself. This would also explain why the Nurses' Health Study observed a 40 percent reduction in the yearly risk of death from all causes among women taking H.R.T. And it would explain why the Nurses' Health Study reported very similar seemingly beneficial effects for antioxidants, vitamins, low-dose aspirin and folic acid, and why these, too, were refuted by clinical trials. It's not necessarily true, but it certainly could be.

While Willett, Stampfer and their colleagues will argue confidently that they can reasonably rule out these other explanations based on everything they now know about their nurses — that they can correct or adjust for compliance and prescriber effects and still see a substantial effect of H.R.T. on heart disease — the skeptics argue that such confidence can never be justified without a clinical trial, at least not when the associations being studied are so small. "You can correct for what you can measure," says Rory Collins, an epidemiologist at Oxford University, "but you can't measure these things with precision so you will tend to under-correct for them. And you can't correct for things that you can't measure."

The investigators for the Nurses' Health Study "tend to believe everything they find," says Barrett-Connor of the University of California, San Diego. Barrett-Connor also studied hormone use and heart disease among a large group of women and observed and published the same association that the Nurses' Health Study did. She simply does not find the causal explanation as easy to accept, considering the plausibility of the alternatives. The latest variation on the therapeutic wisdom on H.R.T. is plausible, she says, but it remains untested. "Now we're back to the place where we're stuck with observational epidemiology," she adds. "I'm back to the place where I doubt everything."

What to Believe?

So how should we respond the next time we're asked to believe that an association implies a cause and effect, that some medication or some facet of our diet or lifestyle is either killing us or making us healthier? We can fall back on several guiding principles, these skeptical epidemiologists say. One is to assume that the first report of an association is incorrect or meaningless, no matter how big that association might be. After all, it's the first claim in any scientific endeavor that is most likely to be wrong. Only after that report is made public will the authors have the opportunity to be informed by their peers of all the many ways that they might have simply misinterpreted what they saw. The regrettable reality, of course, is that it's this first report that is most newsworthy. So be skeptical.

If the association appears consistently in study after study, population after population, but is small — in the range of tens of percent — then doubt it. For the individual, such small associations, even if real, will have only minor effects or no effect on overall health or risk of disease. They can have enormous public-health implications, but they're also small enough to be treated with suspicion until a clinical trial demonstrates their validity.

If the association involves some aspect of human behavior, which is, of course, the case with the great majority of the epidemiology that attracts our attention, then question its validity. If taking a pill, eating a diet or living in proximity to some potentially noxious aspect of the environment is associated with a particular risk of disease, then other factors of socioeconomic status, education, medical care and the whole gamut of healthy-user effects are as well. These will make the association, for all practical purposes, impossible to interpret reliably.

The exception to this rule is unexpected harm, what Avorn calls “bolt from the blue events,” that no one, not the epidemiologists, the subjects or their physicians, could possibly have seen coming — higher rates of vaginal cancer, for example, among the children of women taking the drug DES to prevent miscarriage, or mesothelioma among workers exposed to asbestos. If the subjects are exposing themselves to a particular pill or a vitamin or eating a diet with the goal of promoting health, and, lo and behold, it has no effect or a negative effect — it’s associated with an increased risk of some disorder, rather than a decreased risk — then that’s a bad sign and worthy of our consideration, if not some anxiety. Since healthy-user effects in these cases work toward reducing the association with disease, their failure to do so implies something unexpected is at work.

All of this suggests that the best advice is to keep in mind the law of unintended consequences. The reason clinicians test drugs with randomized trials is to establish whether the hoped-for benefits are real and, if so, whether there are unforeseen side effects that may outweigh the benefits. If the implication of an epidemiologist’s study is that some drug or diet will bring us improved prosperity and health, then wonder about the unforeseen consequences. In these cases, it’s never a bad idea to remain skeptical until somebody spends the time and the money to do a randomized trial and, contrary to much of the history of the endeavor to date, fails to refute it.

Gary Taubes is the author of the forthcoming book “Good Calories, Bad Calories: Challenging the Conventional Wisdom on Diet, Weight Control and Disease.”