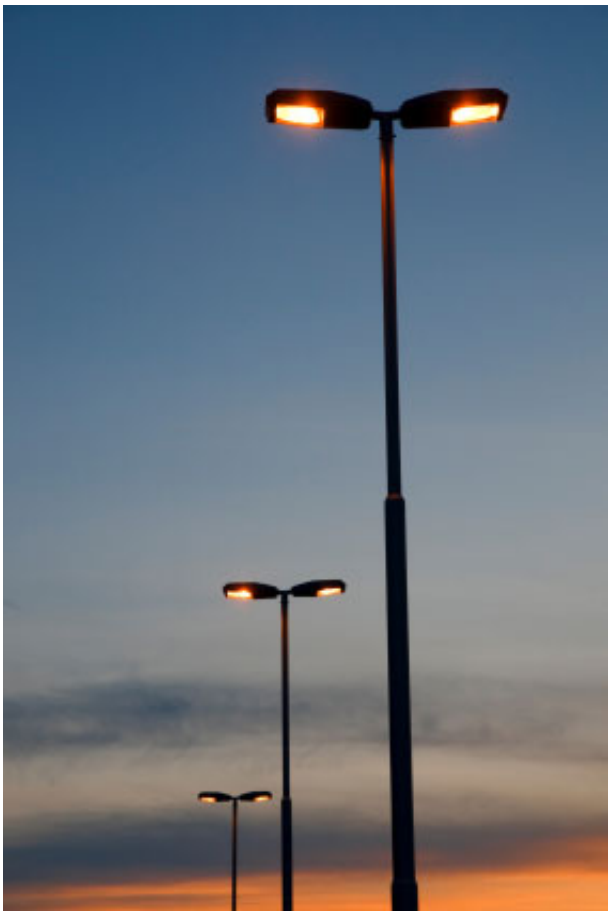
FROM THE JULY-AUGUST 2010 ISSUE

Why Scientific Studies Are So Often Wrong: The Streetlight Effect

Researchers tend to look for answers where the looking is good, rather than where the answers are likely to be hiding.

By David H. Freedman | Friday, December 10, 2010



iStockphoto

A bolt of excitement ran through the field of cardiology in the early 1980s when anti-arrhythmia drugs burst onto the scene. Researchers knew that heart-attack victims with steady heartbeats had the best odds of survival, so a medication that could tamp down irregularities seemed like a no-brainer. The drugs became the standard of care for heart-attack patients and were soon smoothing out heartbeats in intensive care wards across the United States.

But in the early 1990s, cardiologists realized that the drugs were also doing something else: killing about 56,000 heart-attack patients a year. Yes, hearts were beating more regularly on the drugs than off, but their owners were, on average, one-third as likely to pull through. Cardiologists had been so focused on immediately measurable arrhythmias that they had overlooked the longer-term but far more important variable of death.

The fundamental error here is summed up in an old joke scientists love to tell. Late at night, a police officer finds a drunk man crawling around on his hands and knees under a streetlight. The drunk man tells the officer he's looking for his wallet. When the officer asks if he's sure this is where he dropped the wallet, the man replies that he thinks he more likely dropped it across the street. Then why are you looking over here? the befuddled officer asks. Because the light's better here, explains the drunk man.

That fellow is in good company. Many, and possibly most, scientists spend their careers looking for answers where the light is better rather than where the truth is more likely to lie. They don't always have much choice. It is often extremely difficult or even impossible to cleanly measure what is really important, so scientists instead cleanly measure what they can, hoping it turns out to be relevant. After

all, we expect scientists to quantify their observations precisely. As Lord Kelvin put it more than a century ago, "When you can measure what you are speaking about, and express it in numbers, you know something about it."

There is just one little problem. While these surrogate measurements yield clean numbers, they frequently throw off the results, sometimes dramatically so. This "streetlight effect," as I call it in my new book, *Wrong* (Little, Brown), turns up in every field of science, filling research journals with experiments and studies that directly contradict previously published work. It is a tradition that was already well established back in 1915 when an important experiment led by a rather prominent young physicist named Albert Einstein was published. To discover the ratio of magnetic forces to gyroscopic forces on an electron, Einstein had to infer what the electrons in an iron bar were up to based on a minuscule rotation their activity caused the bar to make. His answer was off by a factor of two, as corrected by more careful, but similarly inferential, experiments three years later. (What a loser!)

Physicists have a good excuse for huddling under the streetlight when they are pushing at the limits of human understanding. But the effect also vexes medical research, where you might think great patient data is there for the tabulating. The story of the anti-arrhythmia drugs only hints at the extent of the problem. In 2005, John Ioannidis of the University of Ioannina in Greece examined the 45 most prominent studies published since 1990 in the top medical journals and found that about one-third of them were ultimately refuted. If one were to look at all medical studies, it would be more like two-thirds, he says. And for some kinds of leading-edge studies, like those linking a disease to a specific gene, wrongness infects 90 percent or more.

We should fully expect scientific theories to frequently butt heads and to wind up being disproved sometimes as researchers grope their way toward the truth. That is the scientific process: Generate ideas, test them, discard the flimsy, repeat. In fact, testing ideas is supposed to be the core competence of most scientists. But if tests of the exact same idea routinely generate differing, even opposite, results, then what are we humble nonscientists supposed to believe?

I have spent the past three years examining why expert pronouncements so often turn out to be exaggerated, misleading, or flat-out wrong. There are several very good reasons why that happens, and one of them is that scientists are not as good at making trustworthy measurements as we give them credit for. It's not that they are mostly incompetents and cheats. Well, some of them are: In several confidential surveys spanning different fields, anywhere from 10 to 50 percent of scientists have confessed to perpetrating or being aware of some sort of research misbehavior. And numerous studies have highlighted remarkably lax supervision of research assistants and technicians. A bigger obstacle to reliable research, though, is that scientists often simply cannot get at the things they need to measure.

Examples of how the streetlight effect sends studies off track are ubiquitous. In many cases it is painfully obvious that scientists are stuck with surrogate measures in place of what they really want to quantify. After decades of dueling studies about whether it was an asteroid or volcanic eruptions that did in the dinosaurs, it is apparent that the mineral-deposit evidence is indirect and open to interpretation, even if

the scientists advancing the various claims sound pretty sure of themselves. Astronomers enlist surrogate measures all the time, since there is no way to stick thermometers in stars or to unreel tape measures to other galaxies. Likewise, economists cannot track the individual behaviors of billions of consumers and investors, so they rely on economic indicators and extracts of data.

How reliable are the results? In 1992 a now-classic study by researchers at Harvard and the National Bureau of Economic Research examined papers from a range of economics journals and determined that approximately none of them had conclusively proved anything one way or the other. Given that dismal assessment—and given the great influence of economists on financial institutions and regulation—it's a wonder the global economic infrastructure is not in far worse shape. (Of course, scientific findings that point out the problems with scientific findings are fair game for reanalysis too.)

By far the most familiar and vexing consequences of the streetlight effect show up in those ever-shifting medical findings. Take this straightforward and critical question: Can vitamin D supplements lower the risk of breast, colon, and other cancers? Yes, by as much as 75 percent, several well-publicized studies have concluded over the past decade. No, not at all, several other equally well-publicized studies have concluded. In 2008 alone, around 380 published research articles addressed the link between vitamin D and cancer in one way or another. The ocean of data on the topic is vast, swelling, and teeming with sharp contradictions.

One likely confounding factor is the different ways in which the studies assessed the intake of vitamin D. In fact, some of the studies did not measure intake at all. Researchers simply looked at levels of the vitamin in subjects' blood without tracking whether supplements affected those levels, assuming that both artificially and naturally high levels have the same effect on cancer risk. In some cases researchers looked at blood levels of the vitamin only after a cancer had been diagnosed, instead of measuring the levels before and after. In other cases scientists asked subjects how many vitamin D pills they took but did not look at blood levels. The investigators in at least one widely reported study did not look at vitamin D blood levels or supplement intake at all. They merely estimated blood levels based on the sunniness of the subjects' geographical locations, since sunlight spurs the body to produce vitamin D.

We should expect theories to butt heads as researchers grope their way toward the truth. But if tests of the exact same idea routinely generate different, even opposite, results, what are we supposed to believe?

The point is not that the scientists running these studies screwed up. They were probably doing the best they could with the data they had. We would certainly be a lot more likely to get a straight answer if someone would carefully track pill intake, blood levels, and cancer outcomes in a large population for many years. But such large, clean studies can take years of planning, fund-raising, and lining up patients, plus a decade or more to execute. That is why we first get bombarded by years of weaker studies plagued by the streetlight effect. It sure would be nice if someone would point that out to us when one of those studies makes headlines.

Maybe while we are waiting for that to happen we should pop vitamin D pills just in case, as physicians now commonly recommend. Chances are, that is good advice—unless, of course, the wisdom on vitamin D ends up following the same path as the consensus on aspirin. That consensus long insisted that most people with risk factors for heart disease ought to take a low dose of aspirin every day. But now the prevailing view maintains that unless you have a worrisome history of heart problems, an aspirin regimen is about as likely to hurt you as help you. Oops.

It can take decades to determine whether a drug actually extends lives. That is why researchers more often rely on faster-developing indicators of (apparently) improved health: tumor shrinkage in cancer, lowered blood-sugar levels in diabetes, reduced brain plaque in Alzheimer's, lowered bad cholesterol or elevated good cholesterol in heart disease. Asthma studies alone have looked at nearly 500 different measures of well-being.

The results? We get heavily hyped drugs like Avastin, which shrank tumors without adding significant time to cancer patients' lives (and increased the incidence of heart failure and blood clots to boot); Avandia, which lowered blood sugar in diabetics but raised the average risk of heart attack by 43 percent; torcetrapib, which raised both good cholesterol and death rates; and Flurizan, which reduced brain plaque but failed to slow the cognitive ravages of Alzheimer's disease before trials were finally halted in 2008.

The streetlight effect can also derail study results if scientists do not look at the right subjects. Patient recruitment is an enormous problem in many medical studies, and researchers often end up paying for the participation of students, poor people, drug abusers, the homeless, illegal immigrants, and others who may not adequately represent the population in terms of health or lifestyle. Studies in the 1990s appeared to prove that hormone replacement therapy reduced the risk of heart disease by 50 percent. Then in 2002 a large study seemed to prove that the therapy increased the risk of heart disease by 29 percent. It turns out that a woman's age affects her response to hormone replacement therapy, and the discrepancy arose because the first study looked at somewhat younger women than did the second. The data in both studies were credible; they just did not apply to all women.

Yet another problem is that much of what scientists think they know about human health comes from animal studies. Unfortunately, three-quarters of the drugs that prove safe and effective in animals end up failing in early human trials, sometimes spectacularly. In 2006 the experimental leukemia drug TGN1412 was given to six volunteer human patients. All six of them quickly fell seriously ill with multiple organ trauma, even though the stuff had worked well on rabbits and monkeys at doses up to 500 times as large.

Mice in particular let researchers extract all sorts of exceptionally clean measurements without complaint. Yet it is a well-documented fact that mouse research often translates poorly to human results. Yes, using mice in early drug studies can spare human test subjects from harm, which most people would argue justifies the frequently misleading findings. But mice are also used all the time to obtain easy measurements in harmless lifestyle and behavioral studies. The proposition remains dubious even if the mice are genetically engineered to be more "like" humans in some way. How seriously do you want to

take the advice of a much-hyped 2008 Boston University study declaring that weight lifting can burn more fat than cardio exercise, when the conclusions were based entirely on sedentary mice genetically engineered to have bizarrely large muscles?

Contrary to the proclamations of many scientists, unreliable medical study results do not disappear with large, randomized controlled trials, in which subjects are randomly assigned to a treatment or placebo group. Such trials are more reliable in some ways, but they do not necessarily address the streetlight effect, and they are frequently refuted by other, similar trials. Whatever your take on the healing power of prayer, you have to scratch your head over this: In 1999 a large, randomized controlled trial "proved" that heart surgery patients are more likely to survive if someone they have never met secretly prays for them—and then, seven years later, another randomized trial found that secret prayer very slightly raises the odds that a patient will suffer complications.

The streetlight effect is just one way that research measurements go wrong, and measurement mess-ups are just one of several ways that studies go wrong. Bad measurements are not even the biggest source of wrongness in scientific studies. That honor would have to go to "publication bias"—journals' tendency to eagerly publish the small percentage of studies that produce exciting, surprising, breakthrough results. Of course, the most likely explanation for why one team of researchers comes up with surprising results while several other teams get less-publishable, boring results is that the one team screwed up somewhere. The pernicious effect of this phenomenon on the trustworthiness of study results has been documented at length in scientific journals themselves in several fields. But that's another story.

How are we supposed to cope with all this wrongness? Well, a good start would be to remain skeptical about the great majority of what you find in research journals and pretty much all of the fascinating, news-making findings you read about in the mainstream media, which tends to magnify the problems. (Except you can trust DISCOVER, naturally. And believe me, there is no way *this* article is wrong, either. After all, everything in it is backed by scientific studies.)

Maybe we should just keep in mind what that Einstein fellow—you know, the one who messed up that electron experiment—had to say on the subject: "If we knew what we were doing, it wouldn't be called research, would it?"