

ANDREAS VON BUBNOFF

Numbers Can Lie

FROM *The Los Angeles Times*

SAGITTARIANS are 38 percent more likely to break a leg than people of other star signs — and Leos are 15 percent more likely to suffer from internal bleeding. So says a 2006 Canadian study that looked at the reasons residents of Ontario had unplanned stays in the hospital.

Leos, Sagittarians: there's no need to worry. Even the study's authors don't believe their results. They're illustrating a point: that a scientific approach used in many human studies often leads to findings that are flat-out wrong.

Such studies make headlines every day, and often, as the public knows too well, they contradict each other. One week we hear that pets are good for your health, the next week that they aren't. One month cell-phone use causes brain cancer; the next month it doesn't.

"It's the cure of the week or the killer of the week, the danger of the week," says Barry Kramer, associate director for disease prevention at the National Institutes of Health in Bethesda, Maryland. It's like treating people to an endless regimen of whiplash, he says.

Take the case of just one item: coffee. Drinking two or three cups per day can triple the risk of pancreatic cancer, according to a 1981 study. Not so, concluded a larger follow-up study published in 2001. Coffee reduces the risk of colorectal cancer, found a 1998 study. Not so, according to one published later, in 2005.

"I've seen so many contradictory studies with coffee that I've come to ignore them all," says Donald Berry, chair of the department of biostatistics at the University of Texas M. D. Anderson Cancer Center in Houston.

"What about the man on the street?" asks Stan Young, a statistician at the National Institute of Statistical Sciences in Research Triangle Park, North Carolina. "He reads about coffee causing and not causing cancer — so many contradictory findings he begins to think, 'I don't trust anything these scientists are saying.'"

These critics say the reason this keeps happening is simple: far too many of these epidemiological studies — studies in which the habits and other factors of large populations of people are tracked, sometimes for years — are wrong and should be ignored. In fact, some of these critics say, more than half of all epidemiological studies are incorrect.

The studies can be influential. Often, in response to them, members of the public will go out and dose themselves with this vitamin or that foodstuff. And the studies also influence medical practice; doctors, the critics note, encouraged women to take hormones after menopause long before their effects were tested in randomized clinical trials, the gold standard of medical research.

Some of epidemiology's critics are calling for stricter standards before such studies get reported in medical journals or in the popular press. Young, one of the foremost critics, argues that epidemiological studies are so often wrong that they are coming close to being worthless. "We spend a lot of money and we could make claims just as valid as a random number generator," he says.

Epidemiology's defenders say such criticisms are hugely overblown. They are "quite simplistic and exaggerated," says Meir Stampfer, a professor of epidemiology and nutrition at the Harvard School of Public Health and a professor of medicine at Harvard Medical School.

What's more, some things simply cannot be tested in randomized clinical trials. In certain cases, to do so would be unethical. (Care to assign half the people in a trial to smoke cigarettes?) In other cases, a trial of adequate size and duration — say, to test whether coffee drinking raises or lowers the risk of Parkinson's disease — would have to control the habits of huge numbers of people for decades. That would be not only hugely expensive but also virtually impossible.

Stampfer cites examples of findings of epidemiology that, he says, have stood the test of time: smoking's link to lung cancer, to name the most notable.

Watching for Clues

In epidemiological studies (also called observational studies), scientists observe what's going on; they don't try to change it. From what they observe, they reach conclusions — for example, about the risk of developing a certain disease from being exposed to something in the environment, a lifestyle, or a health intervention.

There are different ways to do this. Cohort studies follow a healthy group of people (with different intakes of, say, coffee) over time and look at who gets a disease. These are considered the strongest type of epidemiological study.

Case-control or retrospective studies examine people with and without a certain disease and compare their prior life — for how much coffee they drank, for example — and see if people who got the disease drank more coffee in their past than those who didn't. Cross-sectional studies compare people's present lifestyle (how much coffee they drink now) with their present health status.

Epidemiological studies have several advantages: they are relatively inexpensive, and they can ethically be done for exposures to factors such as alcohol that are considered harmful, because the people under study chose their exposure themselves.

But epidemiological studies have their minuses, too, some of which are very well known. Suppose a study finds that coffee drinkers are more likely to get a certain disease. That doesn't mean coffee *caused* the disease. Other, perhaps unknown, factors (called "confounders" in the trade) that are unrelated to the coffee may cause it — and if coffee drinkers are more likely to do this other thing, coffee may appear, incorrectly, to be the smoking gun.

A much clearer picture of the role of coffee in disease could be found, in theory, via a randomized clinical trial. Such a study would divide a population in two, put one group on coffee and the other not, then follow both groups for years or decades to see which group got certain diseases and which didn't. The problem, however, is that a randomized trial is very expensive and takes a long time, and it can be difficult to control people's lives for that length of time.

Despite their shortcomings, epidemiological studies are often taken seriously, so much so that they can change medical practice. Such was the case after dozens of epidemiological studies, includ-

ing one large, frequently cited study that came out of Harvard in 1991, had shown that taking estrogen after menopause reduced women's risk of getting cardiovascular disease.

"There was such a belief," even within the medical community, that hormone replacement became part of standard medical practice, says Lisa Schwartz, an associate professor of medicine at Dartmouth Medical School in Hanover, New Hampshire, even in the face of a potential increased risk of breast cancer. In fact, some scientists and doctors said it would be unethical to do a randomized clinical trial to check if the hormone effect was real.

But in the hormone epidemiological studies, women choosing to take hormones may well have been healthier in other ways, Kramer says. And the fact that they were healthier could explain the lower risk of heart disease, not the hormones.

"To get hormone therapy, you have to go to a doctor and have to have insurance," Kramer says. "That means you are in the upper strata of society."

Eventually, a randomized clinical trial was conducted as part of the Women's Health Initiative. Findings published in 2002 not only found no protection to the heart from hormone therapy but actually reported some harm.

Epidemiology's detractors say they have no trouble finding cases other than hormones where frequently cited and sometimes influential epidemiology studies have later turned out to be wrong or exaggerated.

In 1993 Harvard University scientists published two cohort studies reporting that vitamin E protected people from coronary heart disease. One, the Nurses' Health Study, followed more than 87,000 middle-aged female nurses without heart disease for up to eight years. It found that the 20 percent of nurses with the highest vitamin E intake had a 34 percent lower risk of major coronary disease than those with the lowest 20 percent of intake.

The other study followed almost 40,000 male health professionals without heart disease for four years and found a 36 percent lower risk of coronary disease in those men taking more than 60 international units (IU) of vitamin E per day compared with those consuming less than 7.5 IU.

In the three years after these studies appeared, each was cited by other research papers more than four hundred times, according to

John Ioannidis, an epidemiologist at the University of Ioannina School of Medicine in Ioannina, Greece. Vitamin E therapy for heart patients became widespread; a 1997 survey published in the *American Journal of Cardiology* reported that 44 percent of U.S. cardiologists routinely used antioxidants, primarily vitamin E.

The therapy was finally put to the test in a Canadian randomized clinical trial of about 2,500 women and 7,000 men aged fifty-five years or older who were at high risk for cardiovascular events. The findings, reported in 2000, showed that an average daily dose of 400 IU of vitamin E from natural sources for about four and a half years had no effect on cardiovascular disease.

Yet, Schwartz says, seven years after that finding, her patients continue to take vitamin E in the belief that it will protect their hearts. "I am still taking people off vitamin E," she says of her patients, some of whom have heart disease.

Study of Studies

In a provocative 2005 paper, Ioannidis examined the six most frequently cited epidemiological studies published from leading medical journals between 1990 and 2003. He found that in four of the six studies, the findings were later overturned by clinical trials. Vitamin E didn't protect the heart in men or women. Hormone therapy didn't protect the heart in women. Nitric oxide inhalation didn't help patients with respiratory distress syndrome. Another finding turned out later to be exaggerated: taking flavonoids reduced coronary artery disease risk only by 20 percent, not by 68 percent, as originally reported.

The only finding of the six that stood the test of time was a small study that reported that a chemical called all-trans retinoic acid was effective in treating acute promyelocytic leukemia.

The studies that overturned each of these epidemiological findings, Ioannidis says, "caused major waves of surprise when they first appeared, because everybody had believed the observational studies. And then the randomized trials found something completely different."

To be fair, Ioannidis also tested whether the most frequently cited *randomized* studies held up. He found that these had a much better track record. Only nine of thirty-nine oft-cited studies were

later contradicted or turned out to be exaggerated when other randomized studies were done. True, Ioannidis looked at only six studies. But Young says he sees the same trend in his own informal counts of epidemiological claims. "When, in multiple papers, fifteen out of sixteen claims don't replicate, there is a problem," he says.

Belief can be costly, Young adds. For example, one part of the large, randomized Women's Health Initiative study tested the widely held belief — based in large part on epidemiological studies — that a low-fat diet decreases the risk of colorectal cancer, heart disease, or stroke. The findings suggested that there was no effect; "\$415 million later, none of the claims were supported," Young says.

Other scientists, while more cautious than epidemiology's most outspoken detractors, agree that there are many flawed studies. When Kramer first saw Ioannidis's number, "I said to myself, 'It can't be that bad,'" he says. "But I can't prove that it isn't. I know there are a lot of bad studies out there."

Ioannidis, Kramer says, is voicing what many know to be true.

Method in Doubt

Why does this happen?

Young believes there's something fundamentally wrong with the method of observational studies — something that goes way beyond that thorny little issue of confounding factors. It's about another habit of epidemiology that some call data mining.

Most epidemiological studies, according to Young, don't account for the fact that they often check many different things in one study. "They think it is fine to ask many questions of the same data set," Young says. And the more things you check, the more likely it becomes that you'll find something that's statistically significant — just by chance or luck, nothing more.

It's like rolling a pair of dice many times without anyone looking until you get snake eyes and then claiming you'd only rolled it once. Often epidemiological researchers ask dozens, maybe hundreds, of questions in the questionnaires they send to the people they study. They ask so many questions that something eventually is bound to come out positive.

That's where the Canadian star-sign study comes into play, Young says. It was only because the authors deliberately asked a lot of questions — to prove a point — that it was able to come up with significant results for something that couldn't be true.

The study's lead author, statistician Peter Austin of the Institute for Clinical Evaluative Sciences in Toronto, says that once he cleaned up his methodology (by adding a statistical correction for the large number of questions he asked) the association between Leos and internal bleeding and Sagittarians and leg breaks disappeared.

On the Defensive

Many epidemiologists do not agree with the critics' assertion that most epidemiological studies are wrong and that randomized studies are more reliable.

"Randomized studies often contradict one another, as do observational studies," says Harvard's Stampfer, who is an author on both of the frequently cited vitamin E and hormone replacement studies that Ioannidis says were later refuted.

Instead, Stampfer says, the two types of studies often test different things. "It's not an issue here that observational studies got it wrong and randomized trials got it right," he says, referring to the hormone replacement studies. "My view is that [both] were right and they were addressing different questions."

For one thing, the randomized studies on hormone replacement and vitamin E that Ioannidis cited in his 2005 paper looked at different populations than the observational studies they refuted, says Stampfer, who takes vitamin E himself.

In the hormone replacement case, the observational studies looked at women around the age of menopause. The randomized trial looked at women who were mostly well past that age.

In fact, Stampfer says, a recent reanalysis of the Women's Health Initiative data suggested a trend that hormone therapy may be less risky for younger than for older women. The effect was not statistically significant, but, Stampfer says, it's further support for the idea that hormones have different effects depending on when women start taking them.

And the vitamin E studies? The 1993 observational studies fol-

lowed people who didn't have heart disease. The randomized study looked at people with known heart disease who were on many other medications. All those meds could easily override the effect of vitamin E, says Walter Willett, a professor of epidemiology and nutrition at the Harvard School of Public Health, who was a coauthor on both the hormone and the vitamin E epidemiological studies.

And, finally, the low-fat trial from the Women's Health Initiative. It's not surprising, Willett and Stampfer say, that this gold-standard trial failed to find what epidemiology had: that low-fat diets ward off heart disease, colorectal cancer, and stroke; the women in these trials didn't stick to their diets.

"The compliance with the low-fat diet was definitely far lower than anticipated," Willett says, "and probably far worse than even acknowledged in the papers."

Such arguments do not sway epidemiology's detractors. Each time a study doesn't replicate, "they make a specific argument why the studies are different," Young says. He concedes that epidemiology did uncover the truth about the risks of smoking — but only because the effects are so strong. "Even a blind hog occasionally finds an acorn," he says.

Yet epidemiologists warn that discarding results because of a correction for multiple testing may risk missing true and important effects — especially in cases where there's a good biological reason suggesting an effect, such as in studies of drugs that have been shown to work in animal experiments. And setting the bar too high might sometimes be dangerous, says Sander Greenland, a professor of epidemiology and statistics at UCLA. "Do you want to screen for medical side effects with the attitude, 'So what if we miss side effects?'" he asks. "That's deadly. That's ridiculous!"

The debate is unlikely to be resolved any time soon. "If you put five epidemiologists and five statisticians in a room and have this debate," Young says, "and try to get each one to convince the other side, at the end of the day it will still be five to five."