From the Editor: The Replication Crisis is Here

From the Editor: The Replication Crisis is Here

BY MALACHI SHEAHAN III, MD

In 2011, a new research study proved the existence of precognition, also known as Extra-Sensory Perception (ESP). The manuscript was published in the *Journal of Personality and Social Psychology*, a well-regarded peer-review journal with a high-impact factor. The author was Daryl Bem, a star in the field of psychology, best known for his highly influential self-perception theory of attitude.

Bem performed his ESP testing over 10 years, enrolling 1,000 subjects. The trial design consisted of nine separate mini-experiments. In one, participants were told to pick between two curtains on a computer screen, one of which was obscuring a photograph. The location of the photo was then determined at random after the participant made their choice. Subjects who could reliably choose the correct curtain must, therefore, have knowledge of the future. Of the nine similar mini-experiments performed, eight reached statistical significance.

The results of Bem's research forced scientists to make a daunting choice. Should they accept the data and believe the impossible? Or reject it, which would also mean rejecting most of their faith in scientific method? That was the catch; Bem's study design and statistical analysis were nearly flawless.

The report would have fundamentally changed modern science, except for one problem: It could not be reproduced. As other researchers quickly attempted to confirm Bem's amazing findings, most were unsuccessful. These failures spurred wider attempts to duplicate other psychological experiments. Again, most results could not be repeated. An entire scientific field was collapsing. Bem's study was the birth of the "replication crisis" in psychology.

Many scientists and physicians were not surprised by these developments, labeling psychology a pseudo-science. The trouble with dismissing psychology research is that most of the methods and statistics used are universal to science. Would other medical fields be subject to the same dilemma of faith? The answer came quickly. In 2011, Bayer Healthcare attempted to reproduce the results of 47 cancer projects; they were successful in less than 25%. In 2012, Amgen tried to confirm the findings of 53 landmark cancer trials; they succeeded in six.

To understand why our research methods may be fundamentally flawed, it is helpful to look at how they became standard practice in the first place.

Genesis

The birth of modern medical statistics can be traced to a small agricultural research lab in 1920s London, where a young woman preferred to have her tea prepared just so. That woman was Muriel Bristol, an algae biologist. One

afternoon during a break at the lab, a mathematician named Ronald Fisher offered Bristol a cup of tea. She refused the cup, noting that Fisher had poured the milk first and the tea second. Bristol stated that she preferred the milk to be added to the tea, and not the reverse.

Fisher was a brilliant scientist and well versed in thermodynamics. There was absolutely no way that the order of the ingredients mattered. Bristol simply insisted that she could tell the difference. It was here that William Roach, a chemist, intervened. Roach proposed a test, eight cups prepared, four with milk first and four with tea first. Then the cups would be randomly presented to Bristol for tasting.

After sampling, Bristol correctly identified the preparation method of each of the eight cups. The secret, unbeknownst to any of them, is that when milk is poured into tea, more surface area is exposed to the hot water. At temperatures above 160°F, the whey proteins in milk denature and produce a caramel flavor. This process is minimized when the milk is poured first. Regardless of the reason, Bristol could clearly and reliably identify the tea preparation by taste. The experiment had lifelong implications for each of them. Bristol and William Roach were married. Ronald Aylmer (R.A.) Fisher birthed modern statistical theory.

Fisher became obsessed with the results of the trial. Based on the study, he calculated that the odds that Bristol could not discern the tea preparation by taste were one in 70. But what if she had made a mistake? If she had only correctly identified six of eight cups, then the odds would increase to one in four. Therefore, the sample size was too small; having 12 cups would have been a better design to account for error.

Fisher also realized that it was harder to prove something than to disprove it. If one were to hypothesize that Bristol could identify the tea preparation with 100% accuracy, even a 100-cup sample size could not confirm this with absolute certainty. Yet only one incorrect guess would disprove the statement. This was the origin of the null hypothesis. Fisher began to apply his burgeoning theories of study design to his career in agriculture. In their research lab, different fertilizers were compared by putting one on plot A and another on plot B, and so forth. Fisher realized this method was useless because it was “confounded” by the conditions of the plots. What if plot A was more naturally fertile? Fisher then introduced a new concept. The fertilizers would have to be randomized to different plots. Still, that would not be enough to determine if the differences measured were real or random. Fisher’s answer to this problem was a method called analysis of variance, or ANOVA. Now scientists had a statistical tool to differentiate association from causation—to finally figure out what causes what.

Thwarted

Fisher attempted to publish these new methods, but his early career was stifled by Karl Pearson. Pearson was the editor of *Biometrika*, the only statistical journal at that time. Fisher had a combative style and his wars with Pearson severely hindered his academic progress. In one instance, Pearson was trying to design a formula to estimate the effects different variables had on each other when only a small sample size was available. The math was extremely complicated, and Pearson spent years working on individual scenarios. Fisher looked at the problem and, within a week, submitted a solution that was applicable for all cases. Despite it being correct, Pearson initially rejected Fisher’s submission.

Even after publishing the highly influential “Statistical Methods for Research Workers” in 1925, Fisher could not get an academic appointment. Finally, in 1933, Karl Pearson retired and Fisher received an appointment in eugenics at University College London. The position only came with the caveat that he was forbidden to teach statistics; that role was given to Egon Pearson, Karl’s son.

Fisher went on to receive many accolades, including the Copley Medal, the Royal Medal and the presidency of the Royal Statistical Society. In 1952, he was knighted for his contributions by Queen Elizabeth II. Despite his accomplishments, Fisher never held an academic position in statistics.

Fisher’s methods have their limits. In the years before his death, he came out on the wrong side of one of the most important public health debates of the 20th century. In 1944, the British Medical Research Council commissioned Austin Hill to investigate the rising mortality from lung cancer in men. Hill had recently headed a large study on the use of antibiotics for tuberculosis, which became the first randomized controlled trial ever published. Hill and his co-

researcher Richard Doll set out to find the cause of the lung cancer epidemic. Pollution, better detection methods and smoking were the leading suspects. Hill realized that a randomized controlled trial would be impossible in this instance. Instead, he and Doll interviewed 1,400 hospitalized patients, half of whom were ill from lung cancer. They recorded complete medical, social and family histories looking for possible associations. Doll's initial hypothesis was that lung cancer was related to the use of tar on modern roads. About two-thirds of the way through the study, however, Doll was convinced by the data to quit smoking. Hill and Doll published their results in the *British Medical Journal* in 1950. Conclusively, smokers were more likely to have lung cancer—and there seemed to be a dose-dependency related to the number of cigarettes smoked.

Correlation or causation?

While Hill and Doll attempted to match their control group for age, sex and location of residence, other confounding factors were possible. Therefore, a second study was conducted which followed a large group of doctors, some of whom smoked. Mortality data were collected prospectively. The first 36 doctors to die of lung cancer were all smokers. In 1957, the Medical Research Council and the *British Medical Journal* declared that the most reasonable explanation for the results of these trials is that smoking causes lung cancer.

Fisher was now recently retired and an avid pipe smoker. He had the time and motivation for war. Fisher denounced Hill and Doll as spreading anti-tobacco propaganda and suppressing contrary evidence. In a letter to *Nature*, he conceded that smoking and lung cancer were correlated, but he disputed the causation. What if those suffering from the inflammatory effects of lung cancer were using cigarettes to ease their pain? What if a third factor led to both smoking and lung cancer? Was there a shared genetic proclivity to both? Fisher was able to convince many others to his side. Even the president of the American Cancer Society was a skeptic. Fisher died in 1962, never conceding this point. To Fisher, the null hypothesis was never disproven.

Followers of Fisher's methods are often called frequentists. Most of the scientific research performed today is conducted with frequentist techniques. The replication crisis has exposed many potential problems with these methods. Most frequentist approaches require conducting scientific studies in isolation, without incorporating data from prior knowledge.

So Fisher could attack each lung cancer study in isolation while ignoring the mounting preponderance of evidence. Frequentists have also come to rely disproportionately on p values—a tool even Fisher said should be supplementary. Next month in *Vascular Specialist*, we will explore current flaws in study design and statistical methods that could lead to a replication crisis in vascular surgery.

Article Date: Sunday, March 1, 2020
Author: Re-posted from the March 2020 issue of Vascular Specialist
Tags: Vascular Specialist
Article Type: Article