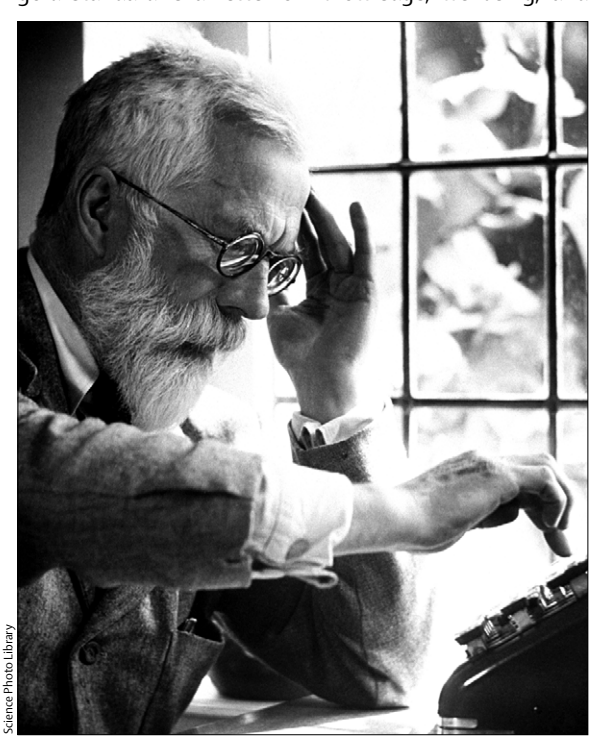
# The Art of Medicine

## The *Validus Medicus* and a new gold standard

Of England's decision to leave the gold standard in 1931, Lord Keynes quipped: "There are few Englishmen who do not rejoice at the breaking of our gold fetters." The gold standard seems to have worked pretty well for big economies up to the First World War. But on the brink of the Depression Keynes thought of gold-based currencies much as a polio-stricken Roosevelt thought of iron leg braces painted gold: they're old, and pretty interesting to look at, sure. But when big movements or fluctuations are required—such as when a big economy changes the backing of its currency, or when a man in iron stockings goes to kick a ball—or when you need to make a fast turn on a dime—the process of converting gold fetters into material welfare is cumbersome and inefficient.

A similar phenomenon can be observed in medical science under a gold standard descending from *Statistical Methods for Research Workers* (1925) by Ronald A Fisher. The gold standard is a set of assumptions about "valid" statistical methods now in use in medicine and other sciences that can't turn on a dime. Since the late 1920s, one incarnation of the gold standard or other has regulated statistical science—from clinical trials on pharmaceuticals to observational studies in psychology and field experiments in poor nations. After Galileo few doubt the need for developing good experimental science. But the statistical gold standard is a fetter on knowledge, wellbeing, and output. It drags down health, raises costs, irritates scientists, and distorts the demand and supply of goods and services by sending incorrect price-quality signals and commodities to the market under the guise of validity and statistical significance. Meantime, good services and commodities are systematically blocked or barely seen through a glass half-cracked—insignificant. There are few scientists who would not rejoice at the breaking of our gold fetters.



Science Photo Library

**Ronald A Fisher (1890–1962)**

The validity of the gold standard is said to consist of three inter-related ideas: randomisation of design; statistical significance; and validity itself—abstractly considered. These are the foundational assumptions underwriting today's rickety standard. As Jennie Freiman, Kenneth Rothman, and others have shown, taking advantage of power, p-value symmetry, and common sense, clinical trials are not as efficacious as they might be, holding cost of treatment effects and other things equal. Too often "certified" experiments on mice and men end prematurely—though the power to detect big effects is discoverable at the time of death. "Student's" tale of randomised control trials (RCTs) is worse. Although practically worshipped, RCTs fail to yield power, precision, and unbiased errors. Finally, statistical significance—the cornerstone of today's gold standard—is not equal to estimation of magnitudes or minimum important difference. Statistical significance at any level does not prove medical, scientific, or commercial importance. We all claim to know this but then we go and do the opposite: we base life decisions on a level of statistical significance.

The assumptions behind the gold standard were instituted ironically on the brink of the Great Depression, by Fisher. A genius at genetics and statistics, Fisher has been described by Richard Dawkins as the greatest biologist since Charles Darwin. But what is his gold standard worth? Not much. The main figure for the neo-Darwinian synthesis in biological thought, Ernst Mayr agrees, was an infallible creationist when it came to statistical and experimental thought. Fisher aspired to conquer the field and he nearly did. His heavy hand still rules. Yet more than a few research workers, classical to Bayesian, would like to break his fetters. He and it, I think, should be replaced by the *Validus Medicus* and a new gold standard—"Lady Platinum", if currency change is good.

Fisher defined validity to mean the theoretically plausible symmetric error distribution that in large samples tends to gather around the mean result. His bell curve is created by imagining numerous replications of the same experiment. Fisher, who had no use for prior information or cost functions, claimed valid results are produced by randomisation which exclusively justifies use of "Student's" *t* test. Errors are valid when observations are independent and experiments are constructed by a table of random numbers.

"Student", our father of statistical Guinness, strongly disagreed. William Sealy Gosset (1876–1937)—better known by his pen name, "Student"—was an Oxford-trained chemist and Fisher's behind-the-scenes adviser and correspondent for 20 years. He pioneered powerful experiments for the laboratory he peacefully lorded over at Guinness brewery for nearly 38 years. "Randomness is a necessary condition for my test", he told Fisher. But "I don't agree with your controlled randomness", "Student" wrote in a letter of October, 1924. "You would want a large lunatic asylum for the operators who are apt to make mistakes enough even at present." Observations are correlated—confounded by ground or other real differences, "Student" found in 1911, requiring balanced layouts.

Few have heard about "validity" in "Student's" powerful 1923, 1936, and 1938 *Biometrika* sense. He gives more of what we want from medicine and other science. His definition of validity derives as science dictates it would from the old Latin roots, *validus* and *valere*, meaning "efficacy", "value", "strength", "robustness". Fisher's definition is, by contrast, nearly valueless to medicine in the Platonic form. "Student" validity is robust and closer to both heavenly and earthly values, measured by beer-significance; by commercial calculation of deliberately chosen small samples; by the power to detect big treatment differences in repeated trials; and by controlling for both real and random sources of fluctuation—as the brewer and healer must. His work at Guinness made the company rich and people happy.

Something about that troubled Fisher, who ran the numbers at Rothamsted Experimental Station. He acquired insufficient knowledge about "Student's" methods before he sterilised them for mass consumption. Unfazed "Student" proved the economic disadvantage of randomisation and fixed rules of significance—impressing Harold Jeffreys, Egon Pearson, and others. Fisher attacked. In March, 1936, Gosset shed the "Student" mask at a meeting of the Royal Statistical Society precisely to show the serious nature of blunders caused by Fisher. He had in his job as Head Experimental Brewer real fish to fry—not philosophy of science. He had to focus on the size of his coefficients; the power of balanced versus random experiments; the cost of observations and new methods, if any; the ease with which any firm could repeat his work on the large-scale. His seminal work inspired Jeffreys' *Theory of Probability* (1961) to the core and it gave the original idea for the Neyman–Pearson research. What Fisher refused to say about "Student" and his economic approach is significant then—however poorly understood.

Take statistical significance. "Student" was dead set against fixed levels of it. "Nearly valueless", he told Egon Pearson in 1937. As Savage noted in *Foundations of Statistics* (1954), statistical significance tells us what to say but not what to do. Still the vague standards of Fisher rule. Think of the he-said, she-said quality of the debate about age as a "significant" factor in mammogram testing. Young women could be forgiven for thinking the need to test is a real coin flip. Take another random example: killing whales. In June, 2005, the Japanese Government increased the limit for the number of whales that may be killed in Antarctica—from 440 whales annually to more than 1000. In the face of international opposition, Deputy Commissioner Akira Nakamae told BBC World News: "We will implement JARPA-2 [the plan for additional killings] according to the schedule, because the sample size is determined in order to get statistically significant results." The Commissioner is standing "Student's" *t* distribution on its head. To "kill more whales" is "to be more significant"—by raising sample size—as if more precise. His backward logic is common in marine biology as much as in field experiments in economics.

Consider the "significance" of damage done in a case involving thousands of human beings, some of them dead. In the early 2000s, quite a few people who took rofecoxib (Vioxx) experienced the wrath of the so-called 5% rule of statistical significance. The clinical trial was published in the *Annals of Internal Medicine* in 2003. The company reported that five patients taking rofecoxib had heart troubles—fatal and not—during the clinical phase. That compared with only one bad result in the control group, "a difference [in bad outcomes] that did not reach statistical significance". After Fisher the erroneous belief is that failing to reach statistical significance is the same as finding no important difference between the two bad outcomes. Not true—Guinness grew rich gambling on the opposite claim. On top of that, investigators discovered they did not report three of eight total bad outcomes—to achieve an insignificant difference, it seems—the error opposite of the one committed by whalers.

There is a holy writ. As Fisher wrote in 1925: "The value for which p=0·05, or 1 in 20, is 1·96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant." In 1926 he said: "Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level."

Let's see what he means: if the p value is exactly 0·05 then the odds that the observed result is real and not random are 0·95/0·05 or 19-to-1. If the p value is 0·12—as it was in an economical but cancelled Illinois Employment Experiment—the odds of a real effect are 0·88/0·12 or 7-to-1. You can fill in the blanks from here—starting with the Great Depression.

What odds should you use when the issue is saving a life, fixing a brain, or feeding the poor? A master brewer said back in 1904 that how he set the odds depends on the importance of the issues at stake. He had a very balanced and valid pint—I mean point—the makings of a new gold standard.

*Stephen T Ziliak*
Department of Economics, Roosevelt University, Chicago, IL 60605, USA

**Further reading**

Goodman S. Toward evidence-based medical statistics, part 2: the Bayes factor. *Ann Intern Med* 1999; **130:** 995–1004.

Gosset WS. "Student's" collected papers. Pearson ES, Wishart J, eds. London: Biometrika Office University College London, 1942.

Jeffreys H. Theory of probability. Oxford: Oxford University Press, 1961.

Rothman KJ, Greenland S, Lash TL. Modern epidemiology. New York: Lippincott, Williams and Wilkins, 2008.

Ziliak ST, McCloskey DN. The cult of statistical significance: how the standard error costs us jobs, justice, and lives. Ann Arbor, MI: University of Michigan Press, 2008.

Ziliak ST. Guinnessometrics: the economic foundation of "Student's" t. *J Econ Perspect* 2008; **22:** 199–216.