When the Alpha is the Omega: *P*-Values, "Substantial Evidence," and the 0.05 Standard at FDA

LEE KENNEDY-SHAFFER*

ABSTRACT

A prominent feature of statistical reasoning for nearly a century, the *p*-value plays an especially vital role in the clinical testing of new drugs. Over the last fifty years, the U.S. Food and Drug Administration (FDA) has relied on *p*-values and significance testing to demonstrate the efficacy of new drugs in the premarket approval process. This article seeks to illuminate the history of this statistic and explain how the statistical significance threshold of 0.05, commonly decried as an arbitrary cutoff, is a useful tool that came to be the cornerstone of FDA decision-making.

INTRODUCTION

The United States Food and Drug Administration (FDA) approved 16 new molecular entities between November 2016 and April 2017, according to the Drugs@FDA database.[1] These new drugs and biologics, whether pills, ointments, or injections, seek to treat conditions as diverse as severe genetic pediatric conditions, dermatitis, chronic kidney disease, constipation, and advanced cancers. The review files for these drugs are a maze of numbers addressing pharmacodynamics, disease incidence and prevalence, doses, results in animal models, and rates of cure or improved symptoms. But one number comes up again and again, regardless of drug class or condition treated: the *p*-value.

This number is used to distill the mountain of information in a New Drug Application (NDA) into understandable and comparable references that describe the overall quantity of evidence. Much maligned and often misinterpreted, the *p*-value plays a central role in guiding decision-making based on statistical evidence in many

*Lee Kennedy-Shaffer, BS, is a PhD student at the Harvard University Graduate School of Arts & Sciences. Department of Biostatistics, Harvard T.H. Chan School of Public Health, 655 Huntington Ave., Building 2, 4th Floor, Boston, MA 02115. Lee_kennedyshaffer@g.harvard.edu. The author's involvement in the writing of this paper was partially in fulfillment of requirements for the course Food and Drug Law at Harvard Law School, taught by Peter Barton Hutt, JD. The author's studies are supported by a grant from the U.S. National Institute of Allergy and Infectious Diseases (5T32AI007358-28). The funder had no role in the design, analysis, preparation, or decision to publish the manuscript. The opinions and analysis in the article are the author's own.

[1] U.S. FOOD & DRUG ADMIN., *Drugs@FDA: FDA Approved Drug Products*, (Jul. 1, 2017) https://www.accessdata.fda.gov/scripts/cder/daf/index.cfm/. The author identified sixteen "Type 1 - New Molecular Entity" approvals in this time frame from the database: Alunbrig, Austedo, Emflaza, Eucrisa, Ingrezza, Kisqali, Parsabiv, Rubraca, Rydapt, Spinraza, Symproic, Trulance, Tymlos, Xadago, Xermelo, Zejula.

disciplines. Nowhere is this role more prominent than in clinical trials, where minute differences in *p*-values can mean the difference between drug approval and failure. Understanding this statistic, and the 0.05 significance level that often accompanies it, requires understanding not only its statistical meaning, but also the history of its use in statistics broadly, and clinical trials specifically. The history of FDA's use of this statistic and this threshold value sheds light on both the outsized role they play in the contemporary drug regulatory regime and the ways in which challenges to this statistic may shape the future of FDA regulation.

I. HISTORY OF RANDOMIZED CONTROLLED TRIALS IN U.S. DRUG REGULATION

In public health and biomedicine, the randomized, blinded, controlled trial is a paradigm of research and often the standard against which other types of evidence are measured.[2] FDA in particular has explicit expectations for the drug development process. Consequently, pharmaceutical and biotechnology companies are very familiar with the three phases of clinical studies, which are largely based on this paradigm.[3] The rise of this system and the specific rules associated with it today, however, have a long history, one that is "neither . . . smooth nor . . . direct" according to historian Harry Marks.[4] In order to understand the role of the *p*-value in the drug approval process, we begin with the source of the data, the clinical trial, and how the trial achieved its scientific and regulatory prominence.

A. The Introduction of the Clinical Trials Paradigm

Early federal drug legislation in the United States focused on prohibiting misbranded and adulterated drugs. In 1938, the Federal Food, Drug, and Cosmetic Act (FDCA) added a prohibition on drugs that were "dangerous to health under the conditions of use prescribed in the labeling thereof."[5] The law also created a premarket notification process whereby a company wishing to market a new drug submitted an application with information about the drug's prescribed use, composition, and safety to the Secretary of Health, Education, and Welfare. Specifically, the application had

---

[2] *See, e.g.*, KENNETH J. ROTHMAN ET AL., MODERN EPIDEMIOLOGY § 6 (3d ed. 2008) (discussing the application of the randomized controlled trial paradigm to other types of epidemiologic evidence in order to make causal claims).

[3] Mark A. Goldberg et al., *Clinical Drug Evaluation and Regulatory Approval*, *in* PRINCIPLES OF PHARMACOLOGY: THE PATHOPHYSIOLOGIC BASIS OF DRUG THERAPY 860, 864–66 (2012).

[4] Harry M. Marks, The Progress of Experiment: Science and Therapeutic Reform in the United States, 1900–1990, at 6 (1997).

[5] David F. Cavers, *The Food, Drug, and Cosmetic Act of 1938: Its Legislative History and Its Substantive Provisions*, 6 LAW & CONTEMP. PROBS. 2, 15 (1939). It is worth noting here that this article will focus on human drugs and biological products regulated via the submission by drug sponsors to FDA of New Drug Applications (for small-molecule drugs) and Biologics License Applications (for biologic products), as specified in Section 505 of the FDCA and regulated now by the Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research, respectively, at FDA. Prior to the assignment of biologics regulation to FDA in 1972, the discussion herein applies primarily to human small-molecule drugs. In 1997, with the FDA Modernization Act, Congress explicitly mandated the harmonization of standards for NDAs and BLAs. Medical devices and animal drugs are regulated under separate frameworks and not discussed here, except for a brief discussion of statistical reviews of efficacy of medical devices, *infra* section V.C. *See* PETER BARTON HUTT ET AL., FOOD AND DRUG L.: CASES AND MATERIALS 135, 1124–31, 1236–38 (4th ed. 2014).

to include "full reports of investigations which have been made to show whether or not such drug is safe for use."[6] These investigations had to "include adequate tests by all methods reasonably applicable" to demonstrate safety.[7] The system did not require active approval, however; if the Secretary failed to reject the application in 60 days, it was automatically approved.[8] For the next 25 years, this system controlled drug approvals in the United States, without any specific reference to drug efficacy.

The mandate for "investigations" and "adequate tests" reflected a shifting paradigm in the U.S. biomedical community. While various randomized or pseudo-randomized experiments had occurred earlier, a recognizable version arose in earnest in the early twentieth century. This modern clinical trial responded to the needs of new biomedical sciences that were developing new therapies at a much faster rate than ever before. With carefully tabulated data available from hospitals and new statistical procedures ready for use, the means to scientifically test drugs became available.[9]

In 1915, Major Greenwood and G. Udny Yule published a paper on cholera and typhoid inoculations that specified the following three specific criteria for valid inference from clinical trials for vaccines: subjects in the inoculated and uninoculated groups must be, "in all material respects, alike"; exposure to the disease must be identical among the inoculated and uninoculated groups; and inoculation and the fact of the disease having occurred must be independent.[10] Other trials used similar designs throughout the 1920s and 1930s, using randomization rather than deliberate balancing in the hopes of fulfilling the first two of those criteria, and the use of placebo controls and blinding of investigators and participants—known as "double-blinding"—in the hopes of fulfilling the third.[11] Major independent research institutes, like the Rockefeller Institute for Medical Research, large hospitals, and medical schools took the lead in conducting these "scientific trials."[12]

In the 1940s and 1950s, large-scale cooperative trials took place on penicillin and other drugs. Replicable methods and established statistical techniques became even more important as the scale of trials grew.[13] In 1948, epidemiologist and biostatistician A. Bradford Hill and the Medical Research Council published results from a large, multi-site study of streptomycin treatment of tuberculosis in Great Britain and, in so doing, set the standard for future efficacy trials. Noting that future investigations of therapeutic agents would "be considered valid only if based on adequately controlled clinical trials," Hill and his colleagues detailed the methodology of their trial in some

---

6 21 U.S.C. § 355(b) (1938).

7 *Id.* § 355(d).

8 Cavers, *supra* note 5, at 40.

9 Harry F. Dowling, *The Emergence of the Cooperative Clinical Trial*, 43 TRANSACTIONS & STUD. C. PHYSICIANS PHILA. 20, 20 (1975).

10 Major Greenwood, Jr. & G. Udny Yule, The Statistics of Anti-typhoid and Anti-cholera Inoculations, and the Interpretation of Such Statistics in General, 8 PROC. ROYAL SOC'Y MED. 113, 115–16 (1915).

11 Abraham M. Lilienfield, *The Fielding H. Garrison Lecture:* Ceteris Paribus: *The Evolution of the Clinical Trial*, 56 BULL. HIST. MED. 1, 14–17 (1982).

12 MARKS, *supra* note 4, at 48–51.

13 *Id.* at 125–26, 132–34, 138–40, 144–48.

detail.[14] The randomization procedures and double blinding, as well as the use of multiple study sites, were particularly discussed. While results were presented in detail, few formal statistical tests were incorporated into this analysis.[15]

Following the publication of the Medical Research Council's trial and a similar streptomycin experiment conducted by the U.S. Public Health Service, controlled randomized experiments became the foundation for the study of pharmaceutical safety and effectiveness. Historian Harry Marks later wrote:

> Since that time, therapeutic reformers have invested controlled randomized experiments with the faith they once had in the integrity and skill of experienced researchers, in the productivity and scientific rigor of cooperative studies, and in the ability of gate-keeping institutions such as the AMA's Council on Pharmacy and Chemistry to transform medical knowledge and practice.[16]

In other words, randomized clinical trials became the gold standard for evaluating drugs among those who wished to put medicine on a truly scientific basis, replacing the myriad forms of evidence clinicians and public health researchers previously considered.[17] Regulators soon followed the trend.

B. The Drug Amendments of 1962 and the "Substantial Evidence" Mandate

Following the thalidomide crisis in Europe in the early 1960s and subsequent lengthy Congressional hearings on the quality of pharmaceutical studies, Congress passed the Drug Amendments of 1962.[18] Also known as the Kefauver-Harris Amendments, these provisions created the first mandate that new drugs be shown to be effective before approval. Specifically, a new basis for refusal of a New Drug Application (NDA) was added to FDCA section 505(d):

> If the Secretary finds . . . that . . . (5) evaluated on the basis of the information submitted to him as part of the application and any other information before him with respect to such drug, there is a lack of substantial evidence that the drug will have the effect it purports or is represented to have under the conditions of use prescribed, recommended, or suggested in the proposed labeling thereof . . . .[19]

---

[14] Streptomycin Treatment of Pulmonary Tuberculosis: A Medical Research Council Investigation, 2 BRIT. MED. J. 769, 769–71 (1948).

[15] *Id*. at 772–82.

[16] Marks, *supra* note 4, at 132.

[17] *See id*. at 2–5 (describing the "political community" of "therapeutic reformers" and how they sought to position medicine as a scientific field).

[18] Robert Temple, *Development of Drug Law, Regulations, and Guidance in the United States*, *in* PRINCIPLES OF PHARMACOLOGY: BASIC CONCEPTS & CLINICAL APPLICATIONS 1643, 1644 (Paul L. Munson et al. eds., 1995). *See also* Jennifer Kulynych, *Will FDA Relinquish the "Gold Standard" for New Drug Approval? Redefining "Substantial Evidence" in the FDA Modernization Act of 1997*, 54 FOOD & DRUG L. J. 132–35 (1999).

[19] Drug Amendments of 1962, Pub. L. No. 87-781, 76 Stat. 781, 781 (codified at 21 U.S.C. § 355(d) (2016)).

That is, a sponsor needed to provide evidence of the drug's efficacy to gain approval. The application also now became a true premarketing affirmative approval process, rather than just an opportunity for the Secretary to reject the application. Moreover, the Secretary was empowered to begin hearings to withdraw approval if new information suggested a lack of substantial evidence of the drug's effectiveness.[20]

The amended section 505(d) then defines "substantial evidence" for this purpose as:

> [E]vidence consisting of adequate and well-controlled investigations, including clinical investigations, by experts qualified by scientific training and experience to evaluate the effectiveness of the drug involved, on the basis of which it could fairly and responsibly be concluded by such experts that the drug will have the effect it purports or is represented to have under the conditions of use prescribed, recommended, or suggested in the labeling or proposed labeling thereof.[21]

This brief definition, appealing largely to expert opinion, makes no mention of statistical principles to be applied, nor does it direct the Secretary or FDA to promulgate any regulations outlining such principles. The only specifics it offers are that the evidence should include "adequate and well-controlled investigations" and these should include "clinical investigations," both plural.[22]

Beginning immediately prior to the passage of the Drug Amendments of 1962, FDA began promulgating its own regulations on clinical trial conduct. Many of these regulations exist through the investigational new drug application (IND) procedure, which is the mechanism by which companies can legally ship experimental drugs in interstate commerce for research purposes prior to FDA marketing authorization.[23] Governed by Section 505(i) of the FDCA, the exemption from the standard rules of interstate drug commerce allows companies to conduct clinical trials, but it also gives FDA the power to regulate trials, a role that can be as significant as the agency desires.[24]

In August 1962, FDA promulgated a Notice of Proposed Rulemaking, which eventually became 21 C.F.R. Part 130, detailing the requirements of the IND process.[25] The final regulations were promulgated in January 1963, after passage of the Kefauver-Harris Amendments. The IND application detailed therein included a

---

[20] *Id.* at 784.

[21] *Id.* at 781.

[22] "Substantial evidence" in other legal contexts has generally been defined as a very low standard of evidence. The Senate Report of the Kefauver-Harris Amendments further suggests that this language was used to mandate evidentiary standards that would, in a legal context, be considered fairly low. Scientifically, "substantial" can be used to denote a wide variety of levels of evidence. *See Drug Efficacy and the 1962 Drug Amendments*, 60 GEO. L. J. 185, 192–95 (1971) (detailing the legislative history of the Kefauver-Harris Amendments' drug efficacy standard and the choice of the "substantial evidence" test); Kulynych, *supra* note 18, at 132–35, 143–47 (detailing the history of the "substantial evidence" standard up to and including the enactment of the FDA Modernization Act of 1997); Jonathan J. Darrow, *Pharmaceutical Efficacy: The Illusory Legal Standard*, 70 WASH. & LEE L. REV. 2073, 2083–88 (2013) (detailing court opinions on the "substantial evidence" standard and legal challenges to the standards employed by FDA).

[23] HUTT ET AL., *supra* note 5, at 674–75.

[24] *Id*. at 674–78.

[25] New Drugs for Investigational Use; exemptions from section 505(a), 27 Fed. Reg. 7990, 7990-92 (Aug. 7, 1962) (to be codified at 21 C.F.R. pt. 130.3).

description of the three phases of trials that were expected: phase 1 on a small number of healthy subjects to determine dose, short-term toxicity, and pharmacological action; phase 2 on a limited number of patients with, or at-risk for, the target condition to determine proof of concept of efficacy; and separate, larger clinical trials in phase 3 to assess drug safety and effectiveness. Initial protocols, including investigator names and qualifications, approximate number of subjects, trial inclusion/exclusion criteria, and trial duration, were expected to be included in this application.[26]

The regulations did not specify strict requirements for the conduct of trials in each phase, in part to "allow flexibility in the design and execution of investigational programs."[27] Meticulous record-keeping, however, was mandated. Additionally, FDA required monitoring of each trial to regularly evaluate safety and effectiveness. Specific language did concur with the statutory language requiring "investigations" (plural), by noting that phase 3 "is conducted by separate groups following the same protocol."[28]

FDA's powers thus arise not from specific language regulating trial design, but from the Commissioner's power to revoke INDs for reasons including an unreasonable plan for clinical investigations and clinical investigations not being conducted according to the submitted plan.[29] Through this rule, then, FDA had the power to terminate the IND exemption from the interstate commerce prohibition, thereby bringing its sponsor's ability to conduct trials to a swift end. The Committee on Public Health of the New York Academy of Medicine, amidst debate over the new regulations in 1962, noted that "it is impossible to lay down one master protocol or procedure for clinical testing" but made clear that the contemporary haphazard state of testing, full of corporate bias and cherry-picking, was not in the best interests of physicians and the public.[30] With these regulations, FDA had carved for itself a massive role, bearing the charge to determine for each drug what constituted "substantial evidence . . . consisting of adequate and well-controlled investigations."[31]

II. THE RISE OF THE *P*-VALUE

The scientific trials FDA demanded, under the 1960s regulations, to support biomedical interventions required a method to summarize the accumulated data. It would be overwhelming to examine a full case report from every trial subject, so physicians and scientists running clinical trials searched for a method to summarize the data. They found such a technique in the papers of statisticians working in various

---

[26] New Drugs: Procedural and Interpretative Regulations; Investigational Use, 28 Fed. Reg. 179, 180 (Dec. 31, 1962). (to be codified at 21 C.F.R. pt. 130.3); New Drugs: Investigational Drugs; Procedure Regarding Biologic Products, 5048, (to be codified at 21 C.F.R. pt. 130.3); New Drugs for Investigational Use; Foreign Shipments; Drugs Used for Diagnosing Disease, 10972-73 (to be codified at 21 C.F.R. pt. 130.3(a)).

[27] New Drugs: Procedural and Interpretative Regulations; Investigational Use, 179, 179 (to be codified 21 C.F.R. pt. 130.3).

[28] *Id.* at 180.

[29] *Id.* at 182.

[30] Committee on Public Health, The Importance of Clinical Testing in Determining the Safety and Efficacy of Drugs, 38 BULL. N.Y. ACAD. MED. 415, 420 (1962).

[31] Drug Amendments of 1962, Pub. L. No. 87-781, 76 Stat. 781, 781 (1962) (codified at 21 U.S.C. § 355(d) (2016)).

fields of biology, and the *p*-value became the standard: one number to summarize the evidence from a clinical trial.

A. The P-Value and Hypothesis Testing

Statistically, the *p*-value serves a very specific function. It is a measure of the compatibility of collected data with a defined scientific hypothesis. In a testing framework, a null hypothesis and an alternative hypothesis are defined. In the health sciences, the null hypothesis is often the absence of some effect, whether of a treatment or intervention, or the absence of a difference between the effects of two treatments. The alternative hypothesis is the opposite of the null hypothesis, generally that some effect or some difference is present. A statistical test is used to determine whether the evidence accords with the null hypothesis.[32]

Within the frequentist framework of statistical inference, there are many ways to formulate statistical tests, and they depend on both the null hypothesis and the data that will be available.[33] A test is generally model-dependent, meaning it relies upon some assumptions about the way in which data are generated. When data are generated according to some probability distribution, properties of that distribution can be used to make an inference about the distribution itself. In general, hypothesis tests involve a test statistic that is a summary of the data; the mean value of some continuous outcome and the number of subjects who experienced some event are two common test statistics. A test then provides ranges of that test statistic for which the null hypothesis will be accepted or rejected.[34]

Within this hypothesis testing framework, the *p*-value, a number between 0 and 1, can be defined in several equivalent ways. The formulation most commonly used in the medical literature defines the *p*-value as the "probability, under the assumption of no effect or no difference, (the *null hypothesis*), of obtaining a result equal to or more extreme than what was actually observed."[35] If an event is only of interest if it is more extreme in the same direction as the observed results (compared to the null hypothesis), then we use only that one-sided probability. More commonly, however, a two-sided probability is calculated that is agnostic to whether the more extreme event is in the same or opposite direction as the observed results. A (one-sided or two-sided) *p*-value is generally then compared to some pre-specified alpha level or significance level. If it is below the alpha level, the null hypothesis is rejected; if it is above the

---

[32] GEORGE CASELLA & ROGER L. BERGER, STATISTICAL INFERENCE 345, 345–46, 364 (1990). *See also, e.g.,* Sander Greenland et al., *Statistical Tests,* P *Values, Confidence Intervals, and Power: a Guide to Misinterpretations*, 31 EUR. J. EPIDEMIOLOGY 337, 338–39 (2016).

[33] The frequentist framework is based on hypothetical repeated sampling of data from some larger population of possible results and assessing the likelihood of the data arising under various scenarios. In contrast, the Bayesian framework considers the test statistic of interest to be a random variable and uses data and prior assumptions to determine the likelihood of various values of that parameter. *See, e.g.,* Jerzy Neyman, *Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability*, 236 PHIL. TRANSACTIONS ROYAL SOC'Y LONDON SERIES A 333, 333–47 (1937) [hereinafter Neyman, *Outline*] (describing the definition of probability used in frequentist methods for estimating parameters and testing hypotheses); Jerzy Neyman, *Frequentist Probability and Frequentist Statistics*, 36 SYNTHESE 97, 113 (1977) (describing, decades later, the frequentist framework that had defined and been shaped by Neyman's work on hypothesis testing); ANDREW GELMAN ET AL., BAYESIAN DATA ANALYSIS, 3, 3–9 (2d ed. 2004) (describing the fundamentals of Bayesian inference).

[34] CASELLA & BERGER, *supra* note 32, at 359.

[35] Steven N. Goodman, *Toward Evidence-Based Medical Statistics. 1: The* P *Value Fallacy*, 130 ANNALS INTERNAL MED. 995, 997 (1999).

alpha level, the null hypothesis is not rejected. One can equivalently define the *p*-value, then, as the value of alpha for which the data would be on the border between rejecting and not rejecting the null hypothesis.[36]

The alpha level is a key part of the testing framework and has caused much controversy. Tests with a fixed alpha level can give rise to two types of errors. A Type I error occurs when the test rejects the null hypothesis despite the null hypothesis being true. A Type II error occurs when the test accepts the null hypothesis despite the null hypothesis being false. Since data are generated from a probabilistic mechanism, chance alone can lead to one of these errors. The alpha level is then the maximum probability of a Type I error; in other words, it is the maximum probability of rejecting the null hypothesis when the null hypothesis is true. The probability of not making a Type II error is called the power of a test, representing a test's ability to detect an effect when an effect exists. A test with a small alpha level is often called a "conservative" test because it is unlikely to reject the null hypothesis when the null hypothesis is true. This often comes with a tradeoff, however; as the test is more likely to accept the null hypothesis when the null hypothesis is false, the power of the test is decreased.[37] For common tests, there is a maximum power that can be achieved for any given significance level.[38]

For determining whether a treatment has an effect on some outcome (or endpoint), tests generally rely on two main features of clinical trials: the effect size and the sample size. The effect size is some measurement of the difference in outcomes between the treatment and control arms. It could be the difference in the proportion of subjects infected by a disease after receiving a vaccine versus without vaccination, for example, or the average difference in serum cholesterol levels after taking a statin compared to taking a placebo pill. The sample size is the number of people enrolled in the trial.

In general, the larger the effect size and the more people in each arm of the trial, the smaller the *p*-value will be. Since the *p*-value is the probability of the observed effect (or a more extreme effect) occurring under the null hypothesis, a smaller *p*-value provides stronger evidence against the null hypothesis and in favor of the alternative hypothesis. So, the smaller the *p*-value is, the harder it is to explain the trial observations simply by appealing to chance variations between the outcomes among the treatment and control subjects. In this hypothesis testing framework, then, the smaller the *p*-value is for a specific trial, the more confident the investigator can be that the drug has an effect.

A confidence interval, which often accompanies a *p*-value in biomedical literature, is a range of estimates of the parameter of interest, i.e., the treatment effect. Under this same frequentist framework, a confidence interval can be calculated as the set of parameter values for the null hypothesis that, with the trial data, would result in a failure to reject the null hypothesis. That is, it is the set of parameter values under which the trial would conclude that the data support the null hypothesis. If the test is conducted with a 0.05 significance level, this is a 95 percent confidence interval. Formally, the frequentist framework does not lend itself to the statement that there is a 95 percent probability of the true effect being within this interval; rather, if the exact same experiment were conducted an infinite number of times, 95 percent of the

---

[36] CASELLA & BERGER, *supra* note 32, at 364.

[37] *Id.* at 358–60.

[38] *Id.* at 365.

intervals generated would include the true effect.[39] The connection between *p*-values, hypothesis test results, and confidence intervals lead to them often being used as surrogates for one another.[40]

B. Early Development of the P-Value

The rise of this number to a central place of importance in a wide variety of disciplines occurred quickly, albeit with only tepid support from statisticians who pioneered its use. Mathematical statistics, including the formalization of assessing uncertainty in data accumulation, is a relatively recent scientific development. Statistics as a field in and of itself arose only in the twentieth century.[41] But it grew out of a long history of using probability models in games of chance, considering the uncertainty in astronomical and geological observations, and in assessing the variation in physical and social processes.[42]

The first known use of a statistic like the *p*-value to assess the likelihood of an observed effect occurring under some null hypothesis came in 1710. While he did not frame it in these terms, the Scottish physician and mathematician John Arbuthnott calculated the probability of male births exceeding female births in London for 82 years in a row. He calculated this probability under the assumption that chance governed the sex of births; that is, that each birth was independent and had equal probability of being a boy or girl. When he found this probability to be exceedingly small (about 1 in 5 septillion), he wrote: "From whence it follows, that it is Art, not Chance, that governs."[43] From this rejection, it is clear that Arbuthnott had decided that this miniscule probability demonstrated sex at birth was not governed by chance in an equally probable way. Following him, physicians and mathematicians studied the regularity of vital statistics (birth and death records) in many different areas with the results usually leading to rejections of chance and acceptance of a divine order.[44]

In 1827, French mathematician Pierre-Simon Laplace, who had already written a major treatise on probability and statistics, used a *p*-value-like statistic and a somewhat more formal hypothesis framework to analyze seasonal barometric pressure measurements. Laplace wrote that a very small value of what would today be the *p*-value "would indicate with a great likelihood that the value of x [the discrepancy between seasons] is not due solely to the anomalies of chance."[45] Finding that very small probability (0.0000015815), Laplace concluded that "the observed discrepancy thus indicates, with an extreme likelihood, a constant cause."[46] From his statements on

---

[39] Neyman, *Outline*, *supra* note 33, at 347–55. *See also id.* at 403–12.

[40] Goodman, *supra* note 35, at 1002.

[41] Stephen Stigler, The History of Statistics: The Measurement of Uncertainty before 1900 1–4 (1986).

[42] *Id.* at 2–5.

[43] John Arbuthnott, An Argument for Divine Providence, Taken from the Constant Regularity Observ'd in the Births of Both Sexes, 27 PHIL. TRANSACTIONS 186, 188–89 (1710).

[44] STIGLER, *supra* note 41, at 226.

[45] 5 PIERRE-SIMON LAPLACE, MECANIQUE CELESTE Supp. 30 (1825 & Supp. 1827) (translating from original French: "[Q]ue la valeur de x n'est pas due aux seules anomalies du hasard.").

[46] *Id.* at Supp. 33 (translating from original French: "L'excès observé indique donc avec une extrême vraisemblance une cause constante . . . .").

other discrepancies that he found not significant, it appears that Laplace implicitly used a 0.01 alpha level in his hypothesis testing.[47]

Around the same time, another Frenchman, Siméon-Denis Poisson, extended Laplace's methods and calculated what we would now call *p*-values and confidence intervals describing the behavior of French juries and whether they were changing due to some cause. In a manuscript in 1837, he noted that a *p*-value of 0.0897 was not strong enough "to support a belief that there has been a notable change in the causes."[48] Poisson used a capital "P" to represent the probability that the observed difference in jury behavior between time periods would be less than or equal to what was observed; that is, his "P" was one minus a modern *p*-value. The notation was likely chosen simply because the value in question is a probability (fortunately, the French *probabilité* also begins with "p").[49] A few pages later, Poisson notes that odds of 200 to one, or a modern *p*-value of about 0.005, are convincing enough to "believe that there was . . . some real anomaly in the votes of juries."[50] He goes on to make a causal claim from this probability statement, attributing the change to the French Revolution of 1830.[51]

Only six years later, Antoine Augustin Cournot examined differences in the proportion of male babies among various population subgroups and calculated a *p*-value, this time using "P" as the modern formulation of the quantity. It represented the a priori chance of the data attaining such a value if the chance-only process (what we now call the null hypothesis) were true.[52] He noted explicitly that "the importance of the deviation δ [between two population or sample means], as given by observation, depends at once on the size of the deviation and on the size of the numbers used," that is, the effect size and the sample size.[53] Cournot explicitly warned of the limits of such probabilistic statements, however, commenting on the importance of the practical meaning of effect sizes and noting that the *p*-value "does not at all measure the chance of truth or of error pertaining to a given judgment."[54] These same concerns are still discussed over 150 years later.

By the late nineteenth century, the concept of considering the probability that observed differences in groups occurred by chance was used in psychology,

---

[47] *Id.* at Supp. 35. *See also* STIGLER, *supra* note 41, at 151.

[48] SIMEON-DENIS POISSON, RECHERCHES SUR LA PROBABILITE DES JUGEMENTS EN MATIERE CRIMINELLE ET EN MATIERE CIVILE 373 (1837) (translating from original French: "[L]a probabilité P . . . ne sont point assez considérables pour qu'on soit bien fondé à croire qu'il y ait eu quelque changement notable dans les causes.").

[49] *See id.* at 372–73. *See also* STIGLER, *supra* note 41, at 189–90.

[50] POISSON, *supra* note 48, at 376–77 (translating from original French: "[O]n peut donc croire qu'il y a eu à cette époque quelque anomalie réelle dans les votes des jurés; et la cause de cette anomalie, qui les a rendus un peu moins sévères, a pu être la Révolution de 1830.").

[51] *Id.* at 377.

[52] ANTOINE AUGUSTIN COURNOT, EXPOSITION DE LA THEORIE DES CHANCES ET DES PROBABILITES 196–97 (1843). *See also* STIGLER, *supra* note 41, at 199–200.

[53] COURNOT, *supra* note 52, at 196 (translating from original French: "Il est évident que l'importance de l'écart δ, comme fait d'observation, dépend à la fois de la grandeur de cet écart et de la grandeur des nombres employés . . . .").

[54] *Id.* at 196–97 (translating from original French: "Quant à la probabilité désignée plus haut par Π, . . . elle ne mesure point la chance de vérité ou d'erreur afférente à un jugement déterminé.").

economics, and other social sciences.[55] In 1885, Francis Ysidro Edgeworth elucidated the significance test in mathematical detail. His procedure took the differences between two populations, and divided them by a "modulus," a function of the sample size and the spread of the individual observations.[56] This procedure is the same one followed today for simple hypothesis testing. Edgeworth used a very conservative test, noting that results due to chance would be "extremely improbable" if they had what we would today call a *p*-value of less than 0.005. The value was seen as a continuous measure of evidence, however, where various such probabilities gave indications of the strength of evidence.[57] Edgeworth further gave numerous examples of situations where one might use this test, ranging from population birth and death rates to economics to the flow of wasps from their nests.[58] This framework, testing significance using a probability model and a null hypothesis, and the *p*-value, even if not referred to as such, had come of age.

C. The P-Value in the Twentieth Century: Application to Randomized Trials

Applying the concept of the probability of extreme results under the null hypothesis to biological settings and, in particular, the randomized trial, came about in the early twentieth century, due largely to the works of Karl Pearson and Ronald A. Fisher. In 1900, Pearson investigated the properties of what is now known as Pearson's chi-squared test of independence, used to analyze tables of outcomes for different populations.[59] Specifically, it often tests the hypothesis of whether the probability of the outcome in population A is different from that in population B.[60] The test statistic, chi-squared (χ^2), follows a specific distribution, from which Pearson calculated probabilities.[61] Already seeing the utility of this statistic, W. Palin Elderton produced an enlarged and improved table of *p*-values (again referred to simply as "P") for χ^2 statistics with given effect sizes and sample sizes.[62]

Building on Pearson's work, William Sealy Gossett—a Guinness brewery employee working on statistical methods for agricultural experimentation and quality control, who published under the pseudonym "Student"—developed an even more general method in 1908. Now known as Student's t-distribution, the distribution arose from Gossett's investigation of the standard deviations, a measure of the spread of data, of random samples. In his landmark 1908 paper describing this method, Gossett analyzed data from a trial of two soporific (sleep-inducing) drugs; he used examples unrelated to his brewery work to avoid revealing his identity and to avoid disclosure

---

55 STIGLER, *supra* note 41, at 260–61, 308–11.

56 F. Y. Edgeworth, *On Methods of Statistics*, J. STAT. SOC'Y LONDON 181, 184–87 (1885).

57 *Id.* at 185.

58 Id. passim.

59 Karl Pearson, On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that It Can Be Reasonably Supposed to Have Arisen from Random Sampling, 50 LONDON EDINBOROUGH & DUBLIN PHIL. MAG. J. SCI. (1900).

60 CASELLA & BERGER, *supra* note 32, at 398.

61 Pearson, *supra* note 59, at 175.

62 W. Palin Elderton, Tables for Testing the Goodness of Fit of Theory to Observation, 1 BIOMETRIKA (1902).

of any trade secrets.[63] He calculated *p*-values for the effect on hours of sleep of each drug separately, and for the performance of one drug compared to the other. In a small sample of 10 patients, Gossett found one-sided *p*-values for a positive effect on sleep of 0.1127 and 0.0026 for drugs 1 and 2, respectively. His standard for evidence was implicitly between these two, as he wrote that "[i]t is then very likely that 1 gives an increase of sleep, but would occasion no surprise if the results were reversed by further experiments" while describing drug 2 as almost certainly effective.[64] He then found a one-sided *p*-value of 0.0015 for the test of the hypothesis that drug 1 and drug 2 had similar effects. Of this he wrote that "odds of this kind make it almost certain that 2 is the better soporific, and in practical life such a high probability is in most matters considered as a certainty."[65] It is important to note that Gossett generally reported either the probability or odds of having a smaller-than-observed effect if the treatment had no impact, so when he speaks of a high probability ("*p*") it corresponds to a low modern *p*-value.[66] Gossett did not fix cutoffs, however, and he warned against fixed levels of significance in tests three decades later, calling them "nearly valueless."[67]

With Gossett's tables of the t-distribution and Elderton's enlarged versions of Pearson's χ^2 tables, methods were in place to use statistics to test hypotheses in experimental trials. The synthesis of these methods and formalization of a general test statistic for hypothesis testing came from the father of modern biostatistics, Ronald Aylmer Fisher. In a 1924 paper and a 1925 monograph, Fisher used the "P" calculations of his forerunners and created a full experimental method in great generality. As is clear from the title, Fisher's book, *Statistical Methods for Research Workers*, was explicitly directed towards practitioners of science, rather than statisticians or mathematicians, as he sought "to put into the hands of research workers, and especially of biologists, the means of applying statistical tests accurately to numerical data."[68] Fisher introduces "P" early in this book, with Tables I and II dedicated to the probability of exceeding various cutoffs under the normal distribution (commonly called a bell curve). It is here that we can trace the beginning of the alpha level of 0.05 as well. Fisher writes:

> The value for which P = .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. Using this criterion we should be led to follow up a negative result only once in 22 trials, even if the statistics are the only guide available. Small effects would still escape notice if the data were insufficiently numerous to bring them out,

---

63 Student (William Sealy Gossett), *The Probable Error of a Mean*, 6 BIOMETRIKA 1, 20 (1908). *See also, e.g.,* Harold Hotelling, *British Statistics and Statisticians Today*, 25 J. AM. STAT. ASS'N 186, 189 (1930) (explaining Gossett's role at Guinness, his efforts to conceal his identity, and the role of statistics in general at the brewery).

64 Student (William Sealy Gossett), *supra* note 63, at 20–21.

65 *Id.* at 21.

66 *Id.* at 20–21.

67 *See* Stephen T. Ziliak, *The* Validus Medicus *and a New Gold Standard*, 376 LANCET 324, 325 (2010).

68 Ronald A. Fisher, Statistical Methods for Research Workers 16 (1925).

> but no lowering of the standard of significance would meet this difficulty.[69]

This passage encapsulates many features of the modern hypothesis test. The 1.96 standard for significance of normally distributed random variables is laid out clearly, as a special case of the use of 0.05 as a reasonable cutoff point for significance. In fact, 0.05 "is convenient" for Fisher precisely because it is the *p*-value associated with two standard deviations from the null within a normal distribution. The normal distribution is especially important because it is a good approximation for many other distributions when sample sizes are large, a result known as the central limit theorem that was proved by Laplace in the early 19th century.[70] Two standard deviations also approximately corresponded to three probable errors, or "quartile distances," of the normal distribution, a measure commonly used during, and prior to, Fisher's era but that eventually was fully replaced by the standard deviation.[71]

It is important to note that Fisher uses a lower value of "P" to indicate more evidence against the null hypothesis, which is the modern formulation of the *p*-value that is most common. Fisher's 1.96 standard is based on the two-sided tests he prefers, although he notes that "P" can be divided by two for a one-sided test.[72] He also clearly states throughout his works that, to make more precise inference, the investigator should collect more data rather than lowering significance thresholds.[73]

Fisher is not entirely precise with his wording interpreting the alpha level, however, especially with regard to the 1 in 22 trials. In fact, of all studies *for which there is no true effect*, an average of 1 in 20 (or 22, depending on the exact cutoff used) will display a significant effect at this threshold. But that is not the same as saying that of all trials we examine, no more than 1 in 20 will be false indications.[74] To obtain this latter probability, one would need to know the proportion of all trials examined for which a true effect existed.

Fisher proceeds to apply his method to other distributions of data and other summary measures that are being tested. For example, he re-defines the value in a chapter on the chi-squared distribution: "P . . . is therefore the probability that χ^2 shall exceed any specified value."[75] Fisher reiterates his 0.05 cutoffs here:

> If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often

---

69 *Id.* at 47.

70 STIGLER, *supra* note 41, at 136–38. The convenience of these standard deviations for the normal were in fact noticed earlier for the special case of the binomial distribution, where each outcome is either a success or failure, and the total result is a count of successes. Abraham De Moivre published the probabilities for being within one, two, and three standard deviations of the mean of a normal distribution in 1733 (translated by the author into English in 1738). His values are very accurate, giving the now-familiar 0.954 for the probability of being within two standard deviations. ABRAHAM DE MOIVRE, THE DOCTRINE OF CHANCES; OR, A METHOD OF CALCULATING THE PROBABILITIES OF EVENTS IN PLAY 238–40 (2d ed. 1738). *See also* STIGLER, *supra* note 41, at 82.

71 FISHER, *supra* note 68, at 47–48.

72 *Id.* at 47.

73 Id.

74 Id.

75 *Id.* at 78.

> be astray if we draw a conventional line at .05, and consider that higher values of χ^2 indicate a real discrepancy.[76]

Fisher uses this formulation, both the 0.05 "significance" threshold and a threshold around 0.01 or 0.02 as "strong evidence," throughout the text in his data analysis examples.

Statistical Methods proved to be groundbreaking both for its harmonization of different distributional methods into the *p*-value significance testing framework and its presentation of new methodology that enabled testing to be performed in a wider range of settings. Fisher introduced many forms of exact tests, which are more accurate in small-sample cases, as well as what is now called the *F*-distribution for the analysis of variances, all using *p*-values.[77] He also discussed randomization principles required for significance tests to be valid for experimental results, work upon which he would later expand.[78]

Fisher's contributions did not end with *Statistical Methods*; in 1935, he wrote a monograph entitled *The Design of Experiments*. In the introduction, Fisher describes the importance of both satisfactory and standardized statistical procedures and elucidates his principles of experimentation.[79] Randomization plays the key role throughout his text, but adjustment for confounding factors (other variables that affect both the probability of receiving a certain treatment and the probability of having the outcome in question), appropriate sample sizes, and determining the mechanisms of chance are treated in detail as well. Notably, Fisher again suggests a 0.05 significance standard (stating that it is "usual and convenient for experimenters to take 5 per cent. as a standard level of significance"), but allows that investigators may wish to specify their own standards based on their purpose and how "exacting" they wish to be.[80]

With a general statistical theory for significance testing and principles for experimental design laid out, Fisher completed his trio of applied biostatistics monographs with *Statistical Tables for Biological, Agricultural, and Medical Research*, written with Frank Yates and published in 1938. In it, the two statisticians presented thirty-four tables of calculated results from common distributions and tests laid out in Fisher's prior works.[81] The book also presented numerous examples of the use of the tables. Most of these focus on the calculation of "P" or construction of a confidence interval from an experiment, demonstrating Fisher's belief in the importance of these measures for a wide variety of statistical work.[82] Equally important, these tables, as they had only limited space for values, almost all exclusively gave values that would be required for determining the 0.05 and 0.01 levels of significance. For the tests of significance for two-by-two contingency tables—tables commonly employed in drug trials showing the number of subjects with each of two outcomes in each of two treatment arms—only these two levels are

---

[76] *Id.* at 79.

[77] *Id.* at 176–210.

[78] *Id.* at 224–29.

[79] Ronald A. Fisher, The Design of Experiments 2–3 (1935).

[80] *Id.* at 15–16.

[81] Ronald A. Fisher & Frank Yates, Statistical Tables for Biological, Agricultural, and Medical Research 25–90 (1938).

[82] *Id.* at 1–22.

given.[83] Future medical statisticians wrote of the importance this had in cementing the status of these values.[84]

How common these standards were in Fisher's time, however, is not entirely clear. Fisher certainly preferred them, and his student L.H.C. Tippett also invoked the 0.05 standard throughout his influential 1931 book *The Methods of Statistics*, calling "the 0.05 level . . . a good compromise" between the two types of error.[85] Tippett also, as is now common, specifically referred to *p*-values less than 0.05 as "*statistically significant*," although he immediately pointed out that his "choice of 0.05 is quite arbitrary" but "in common use."[86] However, Fisher's "statement that it is *usual* for research workers to adopt a 5 per cent significance level in the same context," wrote Lancelot Hogben, "is true only of those who rely on the many rule of thumb manuals expounding Fisher's own test prescriptions."[87] Whether due to a philosophical decision or simply the convenience of Fisher's tables, the level did become a common benchmark over the succeeding decades, referred to even by Fisher's antagonists, Jerzy Neyman and Egon Pearson, and subsequently incorporated as the "most sacred" threshold in papers and textbooks in biology and the social sciences.[88]

With these three books, Fisher crafted a robust framework for significance testing. Randomized experiments of many forms, testing almost any specific parameter, could be translated into test statistics with known distributions. With data from almost any common experiment, an investigator could look up the relevant table and calculate a *p*-value or a confidence interval for the hypothesis or parameter of interest. The stage was thus set for the rise of randomized experiments in the 1940s, as described *supra* section II.A. Perhaps Fisher's greatest contribution was his work to make statistical methods available to investigators without statistical training. But that very work also contributed to the misuse of statistics and overreliance on *p*-values that would later lead to crises of confidence in those methods.

D. Neyman, Pearson, and the Formal Hypothesis Testing Framework

As Fisher was elucidating this view of *p*-values as a continuous measure of evidence, other statisticians were constructing a more formal version of hypothesis testing, one where the results could lead only to a decision to reject or accept a hypothesis. Jerzy Neyman and Egon Pearson, son of Karl Pearson, published their key paper in 1932, in which they described this framework, later known as the Neyman-Pearson method. In their framework, investigators explicitly specify both a null hypothesis and alternative hypothesis, and in the end reject or accept the null

---

[83] *Id.* at 37.

[84] *See, e.g.*, Donald Mainland, *Statistical Ward Rounds—2*, 8 CLINICAL PHARMACOLOGY & THERAPEUTICS (1967).

[85] L. H. C. Tippett, The Methods of Statistics: An Introduction Mainly for Workers in the Biological Sciences 51 (1931).

[86] *Id.* at 48.

[87] Lancelot Hogben, Statistical Theory 495 (1957).

[88] James K. Skipper, Jr. et al., *The Sacredness of .05: A Note Concerning the Uses of Statistical Levels of Significance in Social Science*, 2 AM. SOC. 16, 16 (1967). *See also* Mainland, *supra* note 84.

hypothesis.[89] This decision results in one of three outcomes: a correct determination, a Type I error (incorrectly rejecting the null), or a Type II error (incorrectly accepting the null). The error that is worse "will depend upon the consequences of the error."[90] Critically, the authors recognize that no single decision can be classified as incorrect or correct from purely statistical data, but rather their procedure describes "rules to govern our behavior . . . , in following which we insure that, in the long run of experience, we shall not be too often wrong."[91]

For any data or test statistic arising from data, it is impossible to minimize both types of error. Neyman and Pearson prove, however, that if there is a specified maximum level allowed for probability of a Type I error (denoted epsilon in the paper, but now commonly denoted alpha), then there is a test that minimizes the Type II error for every true parameter.[92] This test is now called the uniformly most powerful test, with power denoting the probability of rejecting the null hypothesis when it is indeed false. The alpha level will depend on the investigator's judgment in weighing the consequences of Type I and Type II errors, but Neyman and Pearson present the familiar 0.05 and 0.01 levels as examples in their article.[93] Neyman and Pearson go on to show that for many common distributions, the test corresponds to "the ordinary test for the significance of a variation in the mean of a sample"[94]; that is, it can be constructed by determining the *p*-value in the way done by Karl Pearson, William Sealy Gossett, and R.A. Fisher. The difference from Fisher's approach is that the *p*-value is no longer a continuous measure of evidence, but rather a test statistic to be compared to a strict cutoff from which a decision is made on the hypotheses.

While Pearson later acknowledged the debt to Fisher's tables and to his stipulation of 0.05 and 0.01 significance levels, a fierce debate raged between Fisher and his two contemporaries about the relative benefits of their frameworks.[95] Fisher's main objections arose from Neyman and Pearson's rejection of the *p*-value as a continuous measure and their emphasis on power. Neyman and Pearson essentially claim that a test with enough power provides evidence *for* the alternative hypothesis rather than simply *against* the null hypothesis. If one of the two must be true, this is a fair statement, but Fisher warned of the many ways in which a null hypothesis can fail, and so did not want to place as much emphasis on the pre-specified alternative.[96] But the Neyman-Pearson framework had a valuable role in decision-making. Fisher himself noted the distinction, describing his methods as a tool for accumulating knowledge rather than for making a final decision.[97]

---

[89] Jerzy Neyman & Egon S. Pearson, *On the Problem of the Most Efficient Tests of Statistical Hypotheses,* 231 PHIL. TRANSACTIONS ROYAL SOC'Y LONDON SERIES A CONTAINING PAPERS OF A MATHEMATICAL OR PHYSICAL CHARACTER, 289, 294–95 (1933).

[90] *Id.* at 296.

[91] *Id.* at 291.

[92] *Id.* at 302.

[93] *Id.* at 303, 305.

[94] *Id.* at 304.

[95] See Erich L. Lehmann, The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?, 88 J. AM. STAT. ASS'N 1242, 1242–44(1993); Johannes Lenhard, Models and Statistical Inference: The Controversy between Fisher and Neyman-Pearson, 57 BRIT. J. PHIL. SCI. 69, 70, 81 (2006).

[96] Lehmann, *supra* note 95, at 1244–45.

[97] Ronald A. Fisher, Statistical Methods and Scientific Inference 99 (1956).

E. The Biomedical Synthesis

In biomedical research, these two approaches—the Fisher "weight of evidence" *p*-value and the Neyman-Pearson formal hypothesis test—have often been combined within a larger frequentist framework. An investigator will specify their null and alternative hypotheses, as well as a pre-specified alpha level (generally, 0.05). She will then calculate a *p*-value from the data and an assumed statistical model. This *p*-value will be compared to the alpha to determine "significance" and the null will be accepted or rejected. The *p*-value will also be presented as a continuous measure, often termed "statistically significant" if it is under 0.05 or "highly statistically significant" if it is under 0.01; some reference to the degree of significance may also be made by comparing the *p*-value to various levels.[98] The rise of other statistical models (for example, the Cox proportional hazards model for survival time) that made use of these frameworks and allowed for calculations of *p*-values and confidence intervals,[99] and the rise of computer software that made calculations of *p*-values easier and more exact than using tables, allowed the *p*-value and the hypothesis or significance testing framework to take precedence in biomedical research.

This synthesis can be seen to some degree in the works of Fisher and of Neyman and Pearson, as well as in the works of statisticians who came soon after them. W. Edwards Deming, in his 1943 book *Statistical Adjustment of Data*, calculated *p*-values—and, in fact, appears to be the first to use the term "*P* value"[100]—and suggested using *p*-values from repeated experiments as measures of the quantum of evidence against the null hypothesis. He also recommended the use of "statistical significance" as an inferential method. But, he warned, "[s]tatistical 'significance' by itself is not a rational basis for action."[101]

Indeed, by the early 1950s, statistics held a prominent place in clinical trials. In the landmark 1948 study of streptomycin by the Medical Research Council, discussed *supra* section II.A, both chi-squared tests and t-tests were used to evaluate the responses to the drug and compare the control and treated groups.[102] At the end of the article, the authors confidently state that "[t]he difference in mortality between the two groups is statistically significant."[103] Interestingly, the authors do not report the calculated *p*-value for any test.

---

[98] Shein-Chung Chow & Jen-Pei Liu, Design and Analysis of Clinical Trials: Concepts and Methodologies 73–75 (2d ed. 2004). *See also* Goodman, *supra* note 35, at 1000–01; Stuart J. Pocock, Clinical Trials: A Practical Approach 204–206 (1983).

[99] *See, e.g.*, David R. Cox, *Regression Models and Life-Tables*, 34 J. ROYAL STAT. SOC'Y SERIES B (METHODOLOGICAL) 187, 187–89 (1972).

[100] W. Edwards Deming, Statistical Adjustment of Data 30 (1943). *See also* H. A. David & A. W. F. Edwards, Annotated Readings in the History of Statistics 223 (2001).

[101] DEMING, *supra* note 100, at 30. This argument was common during the development of these methods and remains so to this day. In its simplest and most common form, it warns of statistical significance supplanting clinical significance or some assessment of effect size altogether as a standard. This is discussed *infra* section V. *See also, e.g.,* FISHER, *supra* note 79, at 194 (ascribing to the experimenter the duty of determining what "observational discrepancy . . . interests him"); STEPHEN T. ZILIAK & DEIRDRE N. MCCLOSKEY, THE CULT OF STATISTICAL SIGNIFICANCE: HOW THE STANDARD ERROR COSTS US JOBS, JUSTICE, AND LIVES 33–42 (2011) (warning of the "sizeless" reliance of statistical inference in several scientific disciplines).

[102] *See* D. D. Reid, *Statistics in Clinical Research*, 52 ANNALS N.Y. ACAD. SCI. 931, 933 (1950).

[103] *Streptomycin Treatment of Pulmonary Tuberculosis: A Medical Research Council Investigation*, *supra* note 15, at 782.

Leading clinical journals soon began to note the importance of such statistical arguments as well. A 1950 editorial in the *Journal of the American Medical Association* (JAMA) posed the question "Are Statistics Necessary?" The answer was an unqualified yes: "If [an investigator] is developing a new therapy, he must know how to set up fourfold tables comparing treated with untreated subjects and must know how to compute the probability that apparently favorable results were accidental."[104] In this language of probability of results due to chance, we see the familiar conceptualization of the *p*-value arise once again. An article in the *Annals of the New York Academy of Sciences* similarly called for quantification of clinical trial results and noted that "[s]tatistical reasoning is needed as soon as that experiment is conceived."[105] Additionally, an article in *JAMA* on the use of controls in medical research presupposed that statistical tests would form the basis of evidence of therapeutic effectiveness, urging clinicians to use randomization and untreated controls as "the basis for statistical comparison" and significance testing.[106]

In the context of drug approvals, both testing frameworks offer advantages. In order to approve a drug, FDA must decide whether the trials provide "substantial evidence that the drug will have the effect it purports or is represented to have,"[107] so a decision-making framework with a strict cutoff is desired. The ability to calculate power is useful to drug sponsors, who have to decide how many patients to enroll in a trial, i.e., how big the sample size will be, in order to demonstrate a true effect exists. But since FDA specifically noted that trial design and evaluation require case-by-case methods, a continuous measure such as Fisher's *p*-value gives a valuable tool to assess the quantity of evidence that a drug has an effect. This tension between the goals of finding as many true effects as possible while not ascribing truth to too many false effects has persisted from its roots in the statistical literature of the 1930s. Today, it survives as the tension that has characterized FDA's drug approval process since 1962, i.e., how to approve all beneficial drugs without approving ineffective drugs.

III. FDA GUIDANCE ON STATISTICAL CONSIDERATIONS IN CLINICAL TRIALS

By 1962, when Congress passed the Kefauver-Harris Amendments, statistical methodologies, including hypothesis testing via the *p*-value, had been combined with the principles of sound experimental design to create an overall structure for clinical drug testing in humans. These principles were not common, much less ubiquitous, in the drug development process, however. Robert Temple, Director of the Office of Medical Policy at the FDA Center for Drug Evaluation and Research, noted that studies submitted in the 1960s often had "no protocol at all. There was almost never a statistical plan."[108] Dr. Louis Lasagna, a prominent pharmacologist at Johns Hopkins University, made a similar point in Senate testimony in 1959, testimony that proved an important precursor to that for the Kefauver-Harris Amendments: "Adequately

---

[104]Editorial, *Are Statistics Necessary?*, 143 J. AM. MED. ASS'N 1260, 1260 (1950).

[105]Reid, *supra* note 102, at 931.

[106]Otho B. Ross, Jr., *Use of Controls in Medical Research*, 145 J. AM. MED. ASS'N 72, 72 (1951).

[107]Pub. L. No. 87-781, 76 Stat. 781 (1962) (codified at 21 U.S.C. § 355(d)).

[108]Robert Temple, *How FDA Currently Makes Decisions on Clinical Studies*, 2 CLINICAL TRIALS 276, 276 (2005).

controlled comparisons of these drugs are almost impossible to find."[109] While he referred specifically to corticosteroids, Dr. Lasagna made clear that his comments generally held true for the drug industry as a whole.

In 1969, FDA sought to overcome this lack of formal scientific and regulatory rigor. The Administration thus began in earnest its role in standardizing clinical trial design protocols. The agency promulgated regulations requiring specific elements in a protocol submitted for an IND. Among these was a "summary of statistical methods used in analysis of the data derived from the subjects."[110] Soon after this, analysis plans became more common and more scientific. Temple noted that all sponsors "came to believe that trials should have a prospectively defined and identified endpoint, a real hypothesis and an actual analytical plan."[111] As a result, FDA began using these statistical tests in decision-making, and the 0.05 standard became enshrined in U.S. drug development.[112]

A. The Rise of the 0.05 Standard in Biomedicine

The usual FDA paradigm traces its roots directly to the regulations implementing the Kefauver-Harris Amendments. The plural "adequate and well-controlled investigations" and subsequent guidance established a standard that two trials following the same protocol should generally be used.[113] Regulations promulgated in 1970 demanded that studies provide "a comparison of the results of treatment or diagnosis with a control in such a fashion as to permit quantitative evaluation."[114] These regulations were somewhat delayed due to prolonged lawsuits between drug manufacturers and FDA over the content of the regulations. Most of these focused on legal principles and had little to do with the substantive definitions in the regulations, but the definitions put forth for "adequate and well-controlled investigations" were considered. In *Pharmaceutical Manufacturers Association v. Richardson*, the U.S. District Court for the District of Delaware found that the requirements, including the "quantitative evaluation" rule and the use of appropriate methods of data analysis, were "minimal requirements for any valid objective study" and thus "not arbitrarily rigid."[115] The regulations, noted the court, "describe broad scientific standards" and still retain flexibility for the sponsor and investigator.[116] The regulations were thus upheld.

---

[109] *Hearings Before the Subcomm. on Antitrust and Monopoly of the Senate Comm. on the Judiciary (pursuant to S. Res 57)*, 86th Cong. 8138–39 (1959) (testimony of Louis Lasagna, M.D., Johns Hopkins University School of Medicine).

[110] 34 Fed. Reg. 14596, 14597 (Sept. 10, 1969).

[111] Temple, *supra* note 108, at 276.

[112] *See* DANIEL P. CARPENTER, REPUTATION AND POWER: ORGANIZATIONAL IMAGE AND PHARMACEUTICAL REGULATION AT THE FDA 269–97 (2010) (detailing FDA's methods of standardizing clinical trial regimes).

[113] Pub. L. No. 87-781, 76 Stat. 781 (1962) (codified at 21 U.S.C. § 355(d)). *See also* 28 Fed. Reg. 179, 180 (Dec. 31, 1962).

[114] 35 Fed. Reg. 7250, 7251 (Apr. 30, 1970).

[115] Pharmaceutical Mfr. Ass'n v. Richardson, 318 F. Supp. 310–11 (D. Del. 1970).

[116] *Id.* at 311. The Court did avoid further defining any of the evidentiary requirements, and refrained from weighing in on the questions of the quantum of evidence and number of required trials. *See Drug Efficacy and the 1962 Drug Amendments*, *supra* note 22.

The regulations, once finally in force, clearly indicated the use of statistical techniques to account for the possibility of random deviations in the presence of no treatment effect. Following common clinical trial practice at the time, this involved the use of significance testing with two-sided tests at the customary significance (or alpha) level of 0.05.[117] While this level for controlling Type I error was not specified in regulations, it was discussed in the biomedical and statistical literature and came to be understood as the customary level, with any deviations from that needing to be pre-specified and defended in the analysis plan.[118] Lancelot Hogben wrote in 1957 that "[c]ontemporary literature of therapeutic and prophylactic trials is an uninterrupted record of Chi Square tests for 2 x 2 tables to test the null hypothesis that there is no treatment difference," referring to one of the main Fisherian methods expounded in the statistician's works.[119] In using these methods, "the overwhelming majority of research workers in the biological field . . . rely largely on rule of thumb procedures set forth in a succession of manuals modeled on *Statistical Methods for Research Workers* by R. A. Fisher," including the 0.05 significance level.[120] Indeed, testing at the 0.05 alpha level became so commonplace in clinical trials that reporting of actual *p*-values was frequently replaced by reporting of only the result of the test, to the chagrin of some biostatisticians.[121]

Reviews of the major drug trials of the time also showed that 0.05 had become accepted practice. In a systematic review of 146 antidepressant drug studies conducted between 1958 and 1972, Jeffrey Morris and Aaron Beck used reported results indicating significant improvement against placebo at the 0.05 significance level.[122] In an analysis of the very large University Group Diabetes Program trial, 0.05 became the standard for significance of a wide variety of outcomes, including the primary outcomes of fatal and nonfatal vascular complications.[123] A review of original research articles in the *New England Journal of Medicine* in 1978 and 1979 found that nearly three-fourths of them used more than descriptive statistics, with *p*-values from hypothesis tests among the most frequent statistical techniques.[124] Major textbooks

---

[117] Bruce A. Barron & Samuel C. Bukantz, *The Evaluation of New Drugs: Current Food and Drug Administration Regulations and Statistical Aspects of Clinical Trials*, 119 ARCHIVES INTERNAL MED. 547, 553 (1967).

[118] *Id.* at 553. *See also* Jerome Cornfield, *Sequential Trials, Sequential Analysis, and the Likelihood Principle*, 20 AM. STAT. 18, 18 (1966); Robert T. O'Neill, *P-Values, Hypothesis Testing, and Reproducibility: An FDA Perspective*, HARV. UNIV. (Apr. 2, 2017), https://catalyst.harvard.edu/pdf/biostatsseminar/O'Neill_Slides.pdf [https://perma.cc/P7WK-U6G6].

[119] Hogben, *supra* note 87, at 497.

[120] *Id.* at 487. Egon Pearson also generally included the 0.05 and 0.01 levels in the tables he published, thus making them generally accessible for the variety of tests investigators wished to conduct in the pre-computer era. *See, e.g.,* Egon S. Pearson, *A Further Development of Tests for Normality*, 22 BIOMETRIKA 239, 240 (1930).

[121] Stuart J. Pocock et al., *Statistical Problems in the Reporting of Clinical Trials: A Survey of Three Medical Journals*, 317 NEW ENGL. J. MED. 426, 431 (1987).

[122] Jeffrey B. Morris & Aaron T. Beck, *The Efficacy of Antidepressant Drugs: A Review of Research (1958 to 1972)*, 30 ARCHIVES GEN. PSYCHOL. 667, 667 (1974).

[123] Alvan R. Feinstein, *Clinical Biostatistics VIII. An Analytic Appraisal of the University Group Diabetes Program (UGDP) Study*, 12 CLINICAL PHARMACOLOGY & THERAPEUTICS (1971).

[124] John D. Emerson & Graham A. Colditz, *Use of Statistical Analysis in* The New England Journal of Medicine, 309 NEW ENGL. J. MED. (1983).

also described the use of *p*-values and significance tests, forming a key part of the education of new clinical investigators.[125]

The appeal by FDA to common practice among medical statisticians is not surprising, especially given the statute's own appeal to "experts qualified by scientific training and experience to evaluate the effectiveness of the drug involved."[126] The practicing trialists, physicians, and statisticians were presumably the experts to whom this statute referred. Significance tests and the 0.05 standard had long since made the leap from the statistical works of Fisher, Neyman, and Pearson, to the medical literature. Dr. Donald Mainland, professor of medical statistics at New York University Medical Center, wrote of the significance level in the *Journal of Clinical Pharmacology and Therapeutics* in 1963.[127] A few years later, he began a series of commentaries in the same journal known as "Statistical Ward Rounds," which addressed the statistical questions of clinical trialists and physicians.[128] His first substantive commentary directly addressed significance testing in drug trials for efficacy, though not always favorably, and was replete with the 0.05 standard.[129] In a similar series entitled "Clinical Biostatistics," Dr. Alvan Feinstein of the Yale School of Medicine continued to use 0.05 as the threshold for significance, even while proposing newer methods of trial design and statistical analysis.[130] As such a key part of the education, work, and publications of trialists, it is natural that *p*-values came to be the accepted form of evidence for the experts making decisions at FDA.

B. Initial FDA Implementation of Statistical Standards

One of the first opportunities for FDA to implement this standard was in the Drug Effectiveness Study, which began in the mid-1960s with the task of determining the effectiveness of drugs on the market prior to the enactment of the Kefauver-Harris Amendments. The National Academy of Sciences/National Research Council undertook the review of thousands of NDAs, classifying each on a six-category scale based on the evidence for the drug's effectiveness.[131] The requirements for evidence were meant to be the same as they would be for new drug applications going forward.[132] Nonetheless, in part because of, as one reviewer put it, a general lack of "statistically valid experimental evidence," this task became quite difficult.[133] The final report of the Drug Efficacy Study is thus more useful in providing intuition on

---

[125] *See, e.g.,* Mainland, *supra* note 84, at 348; POCOCK, *supra* note 98, at 197–206; James H. Ware et al., P *Values*, *in* MEDICAL USES OF STATISTICS 181, 181 (John C. Bailar III & Frederick Mosteller eds., 2d ed. 1992).

[126] Pub. L. No. 87-781, 76 Stat. 781 (1962) (codified at 21 U.S.C. § 355(d)).

[127] Donald Mainland, *Commentary: The Significance of "Nonsignificance"*, 4 CLINICAL PHARMACOLOGY & THERAPEUTICS (1963).

[128] Donald Mainland, *Statistical Ward Rounds—1*, 8 CLINICAL PHARMACOLOGY & THERAPEUTICS 139, 139 (1967).

[129] Mainland, *supra* note 84, at 349–51.

[130] Alvan R. Feinstein, *Clinical Biostatistics V. The Architecture of Clinical Research (concluded),* 11 CLINICAL PHARMACOLOGY & THERAPEUTICS 755, 759 (1970).

[131] *See* HUTT, *supra* note 5, at 776–77.

[132] 35 Fed. Reg. 7250, 7250–51 (Apr. 30, 1970).

[133] NAT'L RESEARCH COUNCIL, DRUG EFFICACY STUDY: FINAL REPORT TO THE COMMISSIONER OF THE FOOD AND DRUG ADMINISTRATION 61–62 (1969).

how evidence would be viewed under the new regime rather than the specific standards employed.

In comments on the process, Drug Efficacy Study reviewers noted many issues with previously conducted clinical trials, encouraged FDA to set forth clear standards, and requested that the agency work with sponsors going forward to ensure appropriate design and analysis.[134] One reviewer specifically brought up statistical principles of design, noting the need for trials to be designed "so that the level of significance of differences between efficacy and spontaneous regression" of disease can be determined.[135] The reviews, publicly available, "identified hundreds, perhaps thousands, of examples of inappropriate, after-the-fact data subsetting . . . , and essentially every other design and statistical 'crime' that could be committed."[136] In the reviews themselves, clear standards were not always set, but the Panel on Antiemitic Drugs (drugs that combat nausea) did explicitly call for statistical analyses "to determine whether observed differences between test and control groups are likely to be caused merely by chance."[137] Given trial practice at the time, this meant significance testing with a pre-specified alpha level. The general principles and issues identified in the Drug Efficacy Study would go on to serve as a basis for FDA guidance in post-1962 NDA considerations as well.

In the late 1970s, statistical tests were showing up in other legal areas as well. In the grand jury discrimination case *Castaneda v. Partida*, the U.S. Supreme Court conducted a hypothesis test and referred to "a general rule" that if the observed data yield a statistic "greater than two or three standard deviations" from the expectation of the null hypothesis, the hypothesis "would be suspect to a social scientist."[138] These standards are equivalent to approximately a 0.05 or 0.01 alpha level for significance testing. The Court supported its use of what are essentially *p*-values in discrimination cases with reference to a 1966 article in the *Harvard Law Review*.[139] In it, Michael Finkelstein calculates *p*-values for several jury discrimination cases and finds them less than 0.05, "the value most commonly used by statisticians."[140] While the statistical test alone does not provide proof of discrimination, writes Finkelstein, it does demonstrate that the proportion of black jury members was not consistent with the racial makeup of the area.[141] These criteria were applied again in the teacher employment discrimination case *Hazelwood School District v. United States* to find statistically significant evidence of discrimination. In this case, however, the Court stated that "these observations are not intended to suggest that precise calculations of statistical significance are necessary in employing statistical proof."[142] While the

---

[134]*Id.* at 92–96.

[135]*Id.* at 96.

[136]Temple, *supra* note 18, at 1656.

[137]NAT'L RESEARCH COUNCIL, *supra* note 133, at 116.

[138]Castaneda v. Partida, 430 U.S. 482, 496 n.17 (1977).

[139]*Id.*

[140]Michael O. Finkelstein, *The Application of Statistical Decision Theory to the Jury Discrimination Cases*, 80 HARV. L. REV. 338, 357–59 (1966).

[141]*Id.* at 359–60.

[142]Hazelwood School District v. United States, 433 U.S. 299, 311 n.17 (1977). The majority and dissenting opinions in *Hazelwood* even differed on whether to use a one-sided or two-sided significance

evidentiary standards in criminal and civil cases are different than those employed by FDA, these cases nonetheless suggest that statistical testing and the 0.05 and 0.01 significance levels had, to some degree, been endorsed by the highest court in the land.[143]

Beginning in the 1970s, FDA expanded its biostatistical corps, which comprised at the time "a few statisticians" with a "modest at best" role in drug review.[144] Statisticians came to contribute "at all levels of review, not only to the review of clinical data and study design, but to the review of" various early-phase studies.[145] This was an ongoing process, however, as Louis Lasagna, who played a major role in formalizing clinical trials and guiding federal drug policy, noted in 1989. He said that even then, a quarter of a century after the passage of the Kefauver-Harris Amendments, FDA was still expanding its role in reviewing clinical trial protocols in the IND submission process.[146]

A high-profile role for the biostatisticians came in 1980, when FDA reviewed a new claim by Ciba-Geigy that its antiplatelet drug, Anturane, was effective in preventing sudden-onset mortality during the first six months after myocardial infarction.[147] The results of the multicenter trial gave *p*-values of 0.058 for cardiac mortality at 24 months and 0.041 for reduction in sudden death over that time, both compared to placebo.[148] The same outcomes were assessed in the period of two through seven months after myocardial infarction as well, both resulting in *p*-values well below 0.05.[149] FDA rejected the claim and published a critique of Ciba-Geigy's results in the *New England Journal of Medicine* explaining the agency's reasoning. FDA reviewers objected to several design features of the study, most notably definitions of outcome events, post hoc exclusions of patients who died during the study due to non-cardiac events, and multiple comparisons. This led the reviewers to state that their "major criticisms of the study are not statistical."[150] Despite this, they calculated adjusted *p*-values accounting for the design flaws and multiple testing, and got values in the 0.12–0.20 range. They finally concluded that the trial was "an insufficient basis for FDA approval."[151] Rather than establishing any specific statistical rules, FDA reviewers demonstrated their commitment to overall principles of study design and control of Type I error. One of the reviewers, Robert Temple, later called the critique "a public tutorial on the analytic

---

test. *See* Paul Meier et al., *What Happened in* Hazelwood*: Statistics, Employment Discrimination, and the 80% Rule*, *in* STATISTICS AND THE LAW 1, 14 (Morris H. DeGroot et al. eds., 1986).

[143]For more on the use of *p*-values and significance testing in the legal system, *see* Meier et al., *supra* note 142, at 6–15; DAVID L. FAIGMAN ET AL., SCIENCE IN THE LAW: STANDARDS, STATISTICS AND RESEARCH ISSUES 188–98 (2002).

[144]Temple, *supra* note 18, at 1655.

[145]*Id.*

[146]Louis Lasagna, *Congress, the FDA, and New Drug Development: Before and After 1962*, 32 PERSP. BIOLOGY & MED. 322, 334 (1989). *See also* CARPENTER, *supra* note 112, at 363 (describing the gradual submission of U.S. pharmaceutical companies to the new FDA protocols).

[147]Robert Temple & Gordon W. Pledger, *The FDA's Critique of the Anturane Reinfarction Trial*, 303 NEW ENGL. J. MED. 1488, 1488 (1980).

[148]Anturane Reinfraction Trial Research Group, *Sulfinpyrazone in the Prevention of Sudden Death after Myocardial Infarction*, 302 NEW ENGL. J. MED. 250, 252–254 (1980).

[149]*Id.*

[150]Temple & Pledger, *supra* note 147, at 1492.

[151]*Id.*

problems that could arise in an otherwise well-conducted study."[152] The influence of statistics at FDA had never been stronger.

C. FDA Guidance and the Formalization of the Two-Trial, 0.05 Standard

Later that decade, FDA published guidance for industry specifying the statistical analyses that would be required for NDA approval. The 1988 guidelines required drug sponsors to detail their statistical analysis plans, including primary outcomes measured and comparisons made. Specifically, FDA requested the use of methods ensuring adequate power and Type I error control, demonstrating that FDA assumed a hypothesis testing procedure with a pre-specified alpha level would be used. The *p*-value is specifically mentioned as a desired feature of the statistical analysis, and the presumption is made that *p*-values would be two-sided unless otherwise specified.[153] While a presumed 0.05 alpha level is not explicitly stated, the presentation in sample tables of 95 percent confidence intervals, which generally coincide with hypothesis tests at the 0.05 alpha level, suggests that a two-sided significance level of 0.05 would be reasonable.[154]

In the 1990s, FDA sought to harmonize its guidance with that of similar agencies in other regions or countries, especially the European Union and Japan. In 1996, the International Conference on Harmonization issued Consolidated Guidance for Industry. The document focused primarily on the design of trials and ensuring ethical treatment of participants. The brief sections on efficacy determinations and statistics, though general, did include the principles of planning statistical tests with pre-specified significance levels and using power calculations to determine appropriate sample sizes.[155]

In 1997, the FDA Modernization Act (FDAMA) amended Section 505 of the FDCA. The law did not specifically adjust the standards for evidence, except in one area detailed *infra* section IV.D. The bigger changes served to encourage FDA to conduct reviews of NDAs more efficiently. For example, the statute mandated that FDA officials meet with drug sponsors "for the purpose of reaching agreement on the design and size of clinical trials intended to form the primary basis of an effectiveness claim."[156] The agreed-upon parameters for the trial would be binding upon the sponsor and the reviewers. While no statistical requirements explicitly entered into the law, the principle of FDA and sponsors agreeing on levels at which to control Type I and Type II errors became codified in the statute.

The next year, FDA released guidance specifically discussing statistical questions in clinical trials. "Statistical Principles for Clinical Trials" did not detail specific procedures or methodologies, but laid out principles to guide the sponsor's statisticians in conducting trial design and analysis. It does, however, specifically address the alpha level and power questions:

---

[152]Temple, *supra* note 18, at 1656.

[153]U.S. FOOD & DRUG ADMIN., GUIDELINE FOR THE FORMAT AND CONTENT OF THE CLINICAL AND STATISTICAL SECTIONS OF AN APPLICATION 68–70 (1988).

[154]*Id.* at 107–08.

[155]U.S. FOOD & DRUG ADMIN./INTL. CONF. ON HARMONIZATION, GUIDANCE FOR INDUSTRY: E6 CLINICAL GOOD PRACTICE: CONSOLIDATED GUIDANCE 41–42 (1996).

[156]Pub. L. No. 105-115, 111 Stat. 2316 (1997) (codified at 21 U.S.C. § 355(b)(5)(B)(i)(I)).

> The treatment difference to be detected may be based on a judgement concerning the minimal effect which has clinical relevance in the management of patients or on a judgement concerning the anticipated effect of the new treatment, where this is larger. Conventionally, the probability of Type I error is set at 5 percent or less or as dictated by any adjustments made necessary for multiplicity considerations; the precise choice may be influenced by the prior plausibility of the hypothesis under test and the desired impact of the results. The probability of Type II error is conventionally set at 10 percent to 20 percent. It is in the sponsor's interest to keep this figure as low as feasible, especially in the case of trials that are difficult or impossible to repeat. Alternative values to the conventional levels of Type I and Type II error may be acceptable or even preferable in some cases.[157]

FDA here explicitly suggested a 0.05 alpha level, but left the door open to other thresholds if justified. In addition, FDA suggested a focus on clinical relevance in determining whether an effect is significant. Later in the document, FDA endorses the general use of two-sided hypothesis tests, in order to match two-sided confidence intervals, unless there is reason to prefer one-sided tests.[158] The 1998 guidance represented the most explicit statement yet of the role of the 0.05 alpha level in FDA's decision-making.

Since the 1998 guidance, two trials, each with a two-sided alpha level of 0.05, has remained the paradigm for FDA approval of drug efficacy. Performing multiple hypothesis tests (either for multiple endpoints, subgroup analyses, or for potentially stopping the trial before its scheduled end) can lead to adjustments of this level, but the goal is generally to ensure the Type I error rate of each trial is less than 0.05. In recent years, guidance has not been as explicit about the 0.05 standard, but a focus on controlling Type I error and summarizing statistical evidence through *p*-values has continued.[159] This continued preference has been borne out in statements by FDA officials, approvals of drug submissions, and the academic literature in both law and biomedicine.[160]

The importance of this standard became clear very quickly, in an FDA decision on United Therapeutics' application for the drug Uniprost (later renamed Remodulin). Submitted in October 2000, the NDA specified an alpha level of 0.049 ("the traditional standard for two confirmatory studies with an adjustment" for one subgroup test).[161] The statistical reviewer found this standard reasonable, and when the *p*-values for the two studies came in at 0.0607 and 0.0550 the reviewer found "no justification for stretching beyond what was specified in the protocol" and was additionally

---

[157]U.S. FOOD & DRUG ADMIN., GUIDANCE FOR INDUSTRY: E9 STATISTICAL PRINCIPLES FOR CLINICAL TRIALS 22 (1998).

[158]*Id.* at 32.

[159]U.S. FOOD & DRUG ADMIN., INTEGRATED SUMMARY OF EFFECTIVENESS: GUIDANCE FOR INDUSTRY 10 (2015).

[160]*See* Joseph W. Cormier, *Advancing FDA's Regulatory Science through Weight of Evidence Evaluations*, 28 J. CONTEMP. HEALTH L. & POL'Y 8–10 (2011); Darrow, *supra* note 22, at 2113–14; Russell Katz, *FDA: Evidentiary Standards for Drug Development and Approval*, 1 NEURORX 307, 310–11 (2004).

[161]*Statistical Review for Uniprost, Application Number: 21-272*, U.S. FOOD & DRUG ADMIN. 3 (2001), https://www.accessdata.fda.gov/drugsatfda_docs/nda/2002/21-272_Remodulin_statr.pdf.

unpersuaded by a *p*-value derived from pooling those studies that was above the threshold for a single study.[162] This review led to a letter by the Director of the Division of Cardio-Renal Drug Products urging the non-approval of Uniprost.[163] In the end, a re-submission focusing on the results of a surrogate endpoint in the studies was approved, conditional upon beginning an additional postmarketing trial.[164]

In addition to showing a relatively strict adherence to the 0.05 threshold, the Uniprost statistical reviewer also offers a more general defense of Type I error control. The reviewer notes FDA's twin goals of keeping ineffective drugs off the market while approving effective drugs and appeals. In language similar to that used by Neyman and Pearson nearly 70 years earlier, he defends the Type I and Type II error control and significance levels traditionally set by FDA as the most appropriate way to strike that balance.[165]

A similar case occurred a few years later with Dendreon's prostate cancer immunotherapy drug, Provenge. The initial submission in 2006 included two trials that had been conducted with a pre-specified significance level of 0.049 (again to adjust for other analyses) for the outcome of disease progression. One of the trials resulted in a *p*-value of 0.052 for this endpoint, which was noted as a failure to meet the primary endpoint by the statistical reviewer.[166] The sponsor attempted to present an efficacy evidence argument based on overall mortality in the trials, which had not been specified as a primary endpoint and did not have a pre-specified significance level in the protocol. Despite impressive *p*-values for these results from each trial, the reviewer noted that these were "post-hoc analyses" and thus it was "difficult to interpret hypothesis test results" for them.[167] He suggested non-approval for efficacy, claiming that "[t]he evidence is not substantial from a statistical perspective," harkening back to the statutory language.[168] FDA did not approve Provenge at that time, in large part due to the statistical shortcomings and lack of Type I error control on the mortality claims.[169] In 2010, after the sponsor conducted another, larger trial with mortality as a pre-specified endpoint and achieved statistically significant results (*p*-value of 0.032), the drug was approved.[170]

---

[162] *Id.* at 18.

[163] Letter from Director, Division of Cardio-Renal Drug Products, U.S. FOOD & DRUG ADMIN., to Director, Office of Drug Evaluation and Review, U.S. FOOD & DRUG ADMIN. (Mar. 9, 2001), https://www.fda.gov/OHRMS/DOCKETS/ac/01/briefing/3775b1_01_Division%20Director's%20memo%20for%20NDA%2021-272.htm.

[164] Letter from Robert Temple, U.S. FOOD & DRUG ADMIN., to Dean Bruce, United Therapeutics Corporation (Feb. 8, 2002), https://www.accessdata.fda.gov/drugsatfda_docs/nda/2002/21-272_Remodulin_Approv.pdf.

[165] *Statistical Review for Uniprost*, *supra* note 161, at 6–7. *Cf.* Neyman & Pearson, *supra* note 89, at 291, 296, 302–05.

[166] *Statistical Review for Provenge, BLA/Serial Number: 125197/0*, U.S. FOOD & DRUG ADMIN. 16, 52 (2007), https://www.fda.gov/BiologicsBloodVaccines/CellularGeneTherapyProducts/ApprovedProducts/ucm210012.htm.

[167] *Id.* at 52.

[168] *Id.*

[169] Letter from Ashok Batra, U.S. FOOD & DRUG ADMIN., to Elizabeth C. Smith, Dendreon Corp. 3 (May 8, 2007), https://www.fda.gov/BiologicsBloodVaccines/CellularGeneTherapyProducts/ApprovedProducts/ucm210012.htm.

[170] *Statistical Review and Evaluation for Provenge, BLA/Serial Number: 125197/0*, U.S. FOOD & DRUG ADMIN. 7, 35–36 (2010), https://www.fda.gov/BiologicsBloodVaccines/CellularGeneTherapy

In all, FDA guidance in the late 1990s made explicit the two-trial, 0.05 standard. Subsequent drug approval decisions reiterated that position. Guidance in May 1998 laid out several reasons for the preference for two Phase 3 trials: (1) reducing the risk of unanticipated, unavoidable biases in any given investigation; (2) reducing the statistical Type I error rate; (3) reducing the possibility of center- or population-specific results leading to wider indications; and (4) reducing the risk of (rare) fraudulent results leading to improper decision-making.[171] This mix of statistical and non-statistical reasons underpins FDA's continued preference for two-trial evidence of efficacy. This preference, however, is not immovable, as is seen through the two other primary routes discussed in that guidance.

D. The Single Trial with Corroborating Evidence Standard

The May 1998 guidance discussed FDA's flexibility in acting on specific cases while illustrating its main paradigms for approval. For decades, though, FDA officials had stated that the two-trial standard, while supported by statute, was not a strict rule, and that a single adequate and well-controlled study could qualify as "substantial evidence."[172] This principle was made explicit with the passage of the FDA Modernization Act in 1997. The statute amended section 505(d) of the FDCA to add that the Secretary of Health and Human Services (or his or her designee, generally the FDA Commissioner), could determine "based on relevant science, that data from one adequate and well-controlled clinical investigation and confirmatory evidence . . . are sufficient to establish effectiveness" and could then use that one study as the basis for approving a drug under the "substantial evidence" standard.[173]

This standard of a single study with corroborating evidence was laid out explicitly in the May 1998 guidance. Various types of corroborating evidence are suggested that may be appropriate, but all of them require a study of the drug with very strong results.[174] FDA specifically stated that any attempt to gain approval of a drug without two trials would leave "little room for study imperfections or contradictory (nonsupporting) information."[175] The remainder of this section of the guidance appealed primarily to biological principles and feasibility rather than statistical arguments.

E. The Single Multi-Center Trial Standard

More compelling statistically, and perhaps more enticing to drug sponsors, was the guidance laid out for a single, multi-center trial standard. As clinical trials grew larger and more complex throughout the decades, the possibility of a single trial providing evidence as convincing as that from two trials grew as well. Prior to the guidance, FDA had made some approvals on the basis of large, single trials, especially if there

---

Products/ApprovedProducts/ucm210012.htm; Letter from Mary A. Malarkey & Celia M. Witten, U.S. Food & Drug Admin., to Elizabeth C. Smith, Dendreon Corp. (Apr. 29, 2010), https://www.fda.gov/BiologicsBloodVaccines/CellularGeneTherapyProducts/ApprovedProducts/ucm210012.htm.

[171] U.S. FOOD & DRUG ADMIN., GUIDANCE FOR INDUSTRY: PROVIDING CLINICAL EVIDENCE OF EFFECTIVENESS FOR HUMAN DRUGS AND BIOLOGICAL PRODUCTS 4–5 (1998).

[172] HUTT ET AL., *supra* note 5, at 725–26.

[173] Pub. L. No. 105-115, 111 Stat. 2313 (1997) (codified at 21 U.S.C. § 355(d)).

[174] U.S. FOOD & DRUG ADMIN., *supra* note 171, at 8–12.

[175] *Id.* at 6.

were specific reasons a second trial would be unfeasible or unethical. In this guidance, FDA made that standard explicit, but still warned that two-trial conclusions are "more secure" than one-trial conclusions.[176]

In the guidance, FDA enumerated five characteristics that may allow a single study to suffice. The study must generally be a multicenter study, have subsets of the study population that show consistent results, have multiple pairwise comparisons showing an effect, and demonstrate an effect on several distinct, important outcomes or endpoints. The final characteristic is that the study shows a "[s]tatistically very persuasive finding." This is interpreted to be a "very low p-value," preferably accompanied by "very sizable treatment effects." While no specific figures are given, the wording suggests that levels of evidence even greater than the traditional alpha level of 0.05, or even 0.01, may be necessary.[177]

The guidance cited two drugs that had succeeded in using this pathway already, timolol for preventing complications after myocardial infarction (MI) and combination streptokinase/aspirin for preventing mortality among patients with suspected MI. Both were cited in particular for their persuasive statistical results.[178] For timolol, the single trial involved 20 hospitals and 1,884 enrolled patients. The primary endpoint of mortality over 33 months after MI was tested and timolol showed a 39.4 percent reduction compared to placebo, with a *p*-value of 0.0003.[179] The streptokinase/aspirin trial (known as ISIS II) had 417 participating hospitals with 17,187 patients randomized into four arms: streptokinase alone, aspirin alone, combination of streptokinase and aspirin, and neither. The combination therapy group had a 42 percent reduction in vascular mortality compared to the placebo, with a *p*-value less than 0.00001. The combination therapy performed better than either therapy individually, with *p*-values less than 0.0001 for the comparison of combination therapy with each individual therapy.[180] In both cases, then, a single, large, multi-center study with substantial effect sizes and very low *p*-values obviated the need for a second confirmatory study in the eyes of FDA.

No concrete standard has been set for what constitutes a "very persuasive" *p*-value. The Uniprost decision discussed *supra* section IV.C referred to a 0.00125 standard as the traditional one-study guidance from the Division of Cardio-Renal Drugs.[181] A presentation on topical microbicides from FDA officials in 2003 suggested that one trial would need a *p*-value less than 0.001 to be considered. They justified this by noting that, absent non-statistical validity considerations, this ensures that the Type I error rate is no greater than the two-trial, 0.05 standard alpha level.[182] One year later,

---

[176] *Id.* at 13.

[177] *Id.* at 13–15.

[178] *Id.* at 12, 15.

[179] Terje Pedersen, *The Norwegian Multicenter Study on Timolol after Myocardial Infarction - Design, Management and Results on Mortality*, 210 J. INTERNAL MED. 235 (1981).

[180] ISIS-2 (Second International Study of Infarct Survival) Collaborative Group, *Randomised Trial of Intravenous Streptokinase, Oral Aspirin, Both, or Neither among 17 187 Cases of Suspected Acute Myocardial Infarction: ISIS-2*, 332 LANCET 349, 354 (1988).

[181] *Statistical Review for Uniprost*, *supra* note 161, at 5–6.

[182] Rafia Bhore & Greg Soon, *Statistical Considerations for Topical Microbicide Phase 2 and 3 Trial Designs*, U.S. FOOD & DRUG ADMIN. 11 (Aug. 20, 2003), https://www.fda.gov/ohrms/dockets/ac/03/slides/3970S1_07_Bhore.ppt [https://perma.cc/Q4X2-8DY8]. *See also Statistical Review for Uniprost*, *supra* note 161, at 7 (using the same reasoning in explaining a one-trial 0.00125 standard). *See also* Zhenming Shun et

another FDA official, also discussing topical microbicides, stated that a single trial, significant at the 0.001 level, would be "persuasive, robust" evidence, while one significant at the 0.01 level would be "acceptable" if the study had other supportive data and good internal consistency.[183] A reviewer of a new drug in 2009 suggested that "less than 0.01" would likely be required and thus rejected "one study with a marginally significant p-value."[184] In perhaps the most definitive statement, Robert Temple remarked in 2005 that "we ordinarily have said that a value in the neighborhood of 0.001 is good enough for a single trial."[185]

The pharmaceutical company Nuvelo and FDA prospectively agreed to the stringent 0.00125 standard for a single trial for approval of the company's thrombolytic agent alfimeprase.[186] The trial, known as SONOMA-2, concluded in 2007. For the primary efficacy endpoint, dissolution of blood clots, the drug achieved a *p*-value of 0.022 against placebo.[187] This result, significant at the traditional 0.05 level but not at the one-trial 0.00125 level, led to Nuvelo withdrawing the NDA submission. Nuvelo's shareholders sued the company, in part alleging that they were misled by the use of a 0.00125 significance level rather than the 0.05 level that has "traditionally been considered convincing evidence by the FDA."[188] The case was settled prior to a decision on the merits by the District Court for the Northern District of California, so the court did not opine specifically on the reasonableness of the 0.00125 significance level.[189]

Notwithstanding some failures, drug sponsors have used the single-trial approval pathway with some frequency since the guidance document, and the two cases cited therein, raised the potential for such approval. A study of new drug approvals at FDA from 2005 to 2015 found that nearly 37 percent of new drugs were approved on the basis of one pivotal efficacy trial (note that this may include drugs that fell under the pathway described *supra* section IV.D as well as the purely single-trial standard). The proportion was highest among cancer drugs, with over 80 percent of new cancer drugs approved on the basis of only one trial.[190] In an address to Congress shortly before stepping down from her role, FDA Commissioner Margaret Hamburg, in defending

---

al., *Statistical Consideration of the Strategy for Demonstrating Clinical Evidence of Effectiveness—One Larger vs Two Smaller Pivotal Studies*, 24 STAT. MED. 1622–28 (2005) (detailing the underlying statistics for combining evidence from different studies and controlling overall Type I error).

[183]Teresa C. Wu, *Clinical Development of Topical Microbicides: U.S. Regulatory Perspective*, U.S. FOOD & DRUG ADMIN., 13 (2004), https://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/HowDrugsareDevelopedandApproved/ApprovalApplications/InvestigationalNewDrugINDApplication/Overview/UCM166921.pdf [https://perma.cc/AAE3-AFRF].

[184]*Statistical Review and Evaluation for Bystolic, NDA/Serial Number: 22-742*, U.S. FOOD & DRUG ADMIN. 3 (2009), https://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/CardiovascularandRenalDrugsAdvisoryCommittee/UCM196558.pdf [https://perma.cc/7GXV-24T7].

[185]Temple, *supra* note 108.

[186]In re Nuvelo, Inc. Securities Litigation 688 F. Supp. 2d 1218, 1221 (N.D. Cal. 2009).

[187]Stephan Moll et al., *Safety and Efficacy of Alfimeprase in Subjects with Occluded Central Venous Access Devices: The SONOMA-2 Study* 110(11) BLOOD 552A–53A (2007).

[188]In re Nuvelo, Inc. Securities Litigation 688 F. Supp. 2d 1218, 1221 (N.D. Cal. 2009).

[189]In re Nuvelo, Inc. Securities Litigation No. 3:07 (N.D. Cal. July 13, 2012) (order cancelling hearing on lead plaintiffs' unopposed motion for distribution of settlement fund).

[190]Nicholas S. Downing et al., *Clinical Trial Evidence Supporting FDA Approval of Novel Therapeutic Agents, 2005–2012*, 311 J. AM. MED. ASS'N 375 (2014).

the speed, rigor, and flexibility of the agency's approval process, noted similarly that one-third of new drugs were approved on the basis of single clinical trials.[191]

Between the FDA Modernization Act and the guidance documents of the late 1990s, FDA set down fairly clear expectations regarding statistical analyses to be included in drug approval submissions. While not an explicit or hard-and-fast rule, the 0.05 alpha standard has remained, generally implicitly, in these guidelines and has even given rise to a general 0.001 alpha standard for a single-trial approval. Fisher's appeal to a "convenient" and "customary" level in the 1920s and 1930s has thus survived over eight decades to inform policy today. But just as Fisher's hypothesis testing framework faced challenges in his own time, FDA's standards and the *p*-value as a whole have faced challenges in recent years.

IV. STATISTICS AT FDA: CONTEMPORARY CHALLENGES AND FUTURE DIRECTIONS

Ever since the initial disagreements between Neyman and Pearson and Fisher, the *p*-value and the 0.05 alpha level standard have caused controversy. Statisticians have debated the proper use and interpretation of the *p*-value, to the point of questioning whether it belongs in statistical reasoning at all. Policymakers and FDA stakeholders have questioned whether the roles of the *p*-value and the 0.05 standard in drug approvals have been appropriate. And there has been no shortage of alternatives presented. Even as the field of statistics has changed and FDA's standards have adjusted to new trial designs and statistical analysis plans, the *p*-value has not lost its influence, or controversy, at the agency.

A. Challenges to the P-Value Paradigm

The challenge to Fisher's interpretations by Jerzy Neyman and Egon Pearson, described *supra* section III.D, was only the first challenge to the regime of significance testing via the *p*-value. In the 1960s, as FDA was incorporating significance testing into its new drug efficacy pre-approval regime, prominent psychologists pointedly questioned the appropriate role of the *p*-value in their field. In 1960, William Rozeboom wrote in the *Psychological Bulletin* of the failings of the *p*-value and the significance testing regime. While many of his arguments focused on his philosophical disdain for the idea of accepting or rejecting a scientific hypothesis outright, he also addressed the more mathematical question of the significance level to be used. "There is no reason (at least provided by the method)," Rozeboom wrote, "why the point of statistical 'significance' should be set at the 95% level, rather than, say the 94% or 96% level. Nor does the fact that we sometimes select a 99% level of significance, rather than the usual 95% level, mitigate this objection—one is as arbitrary as the other."[192] He comes back to this point later in the article, questioning "what scientist in his right mind would ever feel that there is an appreciable difference between the interpretative significance of data, say, for which one-tailed $p = .04$ and that of data

[191] Statement of Margaret A. Hamburg, U.S. FOOD & DRUG ADMIN., to Committee on Health, Education, Labor and Pensions, U.S. Sen. (Mar. 10, 2015), https://www.fda.gov/newsevents/testimony/ucm437481.htm [https://perma.cc/MLP9-LA8J].

[192] William W. Rozeboom, *The Fallacy of the Null-Hypothesis Significance Test*, 57 PSYCHOL. BULL. 423 (1960).

for which $p = .06$, even though the point of 'significance' has been set at $p = .05$?"[193] Rozeboom suggests alternatives that have now become common in journals, and are becoming more common in FDA reviews: the confidence interval and Bayesian inference, both discussed in more detail *infra* section V.B.

Not long after Rozeboom, in his own words, "vigorously excoriated" the significance testing procedure,[194] other psychologists continued the argument. David Bakan, in the same publication in 1966, lamented "a kind of essential mindlessness in the conduct of research" when investigators focus, to the exclusion of other inference, on *p*-values and significance testing.[195] He agreed with Rozeboom's prescription to focus more on effect sizes, confidence intervals, and Bayesian methods. In a far-reaching paper in *Philosophy of Science* in 1967, Paul Meehl contrasted the use of statistical tests in physics, wherein they serve to quantify the uncertainty of a numerical estimate, with their use in psychology, wherein they serve to accept or reject a null hypothesis.[196] All of these papers, along with others along the same lines, point out that a significance testing framework is more appropriate for a situation where a decision one way or another must be made than for the general principle of scientific inference. They also, however, question the model of significance testing more fundamentally, pointing out frequent misinterpretations of its results and the methodologically unjustified but "widespread adoption of the probabilities .01 or .05 as the allowable theoretical frequency of Type I errors" in the biological and social sciences.[197]

While psychologists and social scientists lamented the strict significance tests, they often faced different questions than biomedical researchers, especially those analyzing clinical trials. One sociologist wryly noted that "usually the only real decision facing a social scientist is whether to publish or suppress his findings,"[198] a sharp contrast with the clear-cut decision to approve or reject faced by FDA. In addition, biologists working under controlled conditions in randomized experiments have more control over design and thus may accept different thresholds and stricter significance tests than social scientists. The sociologist Sanford Labovitz wrote that "[u]nder such highly controlled conditions [of agricultural experiments] Fisher seemed justified in using the larger error rate of .05 instead of .01 or lower."[199]

Nonetheless, in biology and medicine, critiques began to appear along similar lines as in the social sciences. As early as 1951, Fisher's collaborator Frank Yates wrote of his concerns about the use of tests with strict levels of significance: "scientific workers have often regarded the execution of a test of significance on an experiment as the

---

[193] *Id.* at 424.

[194] *Id*. at 428.

[195] David Bakan, *The Test of Significance in Psychological Research*, 66 PSYCHOL. BULL. 436 (1966).

[196] Paul E. Meehl, *Theory-Testing in Psychology and Physics: A Methodological Paradox*, 34 PHIL. SCI. (1967).

[197] *Id.* at 106.

[198] Skipper et al., *supra* note 88, at 158.

[199] Sanford Labovitz, *Criteria for Selecting a Significance Level: A Note on the Sacredness of .05*, THE SIGNIFICANCE TEST CONTROVERSY—A READER, 166, 169 (Denton E. Morrison & Ramon E. Henkel eds., 1970).

ultimate objective."[200] In a speech delivered to the International Biometric Society in 1969, the outgoing British Regional President J.G. Skellam presented a defense of confidence intervals and some support for Bayesian ideas, warning that doctrinaire use of Fisherian significance tests might "exercise their own unintentional brand of tyranny over other ways of thinking."[201]

The objections grew in the 1980s as the leading medical journals began to call for more discussion of point estimates and confidence intervals, in addition to *p*-values and significance testing. In 1988, in promoting these approaches, the statistical adviser to the *British Heart Journal* provocatively titled his editorial "The end of the *p* value?"[202] Other prominent medical and epidemiologic researchers, including Kenneth Rothman, Richard Simon, and Steven Goodman, have also written along the same lines.[203] Rothman, in an editorial for *Annals of Internal Medicine* in 1986, wrote that "[t]esting for statistical significance continues today not on its merits as a methodological tool but on the momentum of tradition."[204]

Common complaints about the *p*-value and significance testing have rested not only on methodological grounds, but also on the improper interpretation of results. The most common, and perhaps most reviled among statisticians, misinterpretation is the changing of the conditional probability.[205] In this falsehood, the *p*-value is taken to be the probability that the null hypothesis is true given the data. In fact, the *p*-value is the probability that the data would have occurred if the null hypothesis were true, and these two probabilities are very rarely the same. Another common error is ascribing clinical significance to a statistically significant result. A very large trial may achieve statistical significance at the 0.05 level even without any clinically meaningful difference in outcomes. And the overuse of testing procedures, without adjusting the significance level, or choosing tests based on seeing the data, known as "*p*-hacking," has come under fire recently.[206]

The controversies about *p*-values came to a head in the 2010s, with some journals beginning to discourage or ban outright the use of the *p*-value in their pages.[207] In response to this, the American Statistical Association issued a wide-ranging statement on statistical significance and *p*-values in 2016.[208] With contributions and complementary articles by statisticians with a range of philosophical frameworks and preferred methodological techniques, the statement is hardly definitive or prescriptive.

---

[200]Frank Yates, *The Influence of* Statistical Methods for Research Workers *on the Development of the Science of Statistics*, 46 J. AM. STAT. ASS'N 33 (1951).

[201]J. G. Skellam, *Models, Inference, and Strategy*, 25 BIOMETRICS 474 (1969).

[202]Stephen J. W. Evans et al., *The End of the P Value?*, 60 BRIT. HEART J. (1988).

[203]*See, e.g.,* Steven N. Goodman, *A Comment on Replication, P-Values, and Evidence*, 11 STAT. MED. (1992); Goodman, *supra* note 35; Richard Simon, *Confidence Intervals for Reporting Results of Clinical Trials*, 105 ANNALS INTERNAL MED. (1986); Kenneth J. Rothman, *Significance Questing*, 105 ANNALS INTERNAL MED. (1986).

[204]Rothman, *supra* note 203, at 447.

[205]*See, e.g.,* Goodman, *supra* note 35.

[206]*See, e.g.,* Greenland et al., *supra* note 32, at 342–43, 346; Regina Nuzzo, *Statistical Errors*, 506 NATURE (2014).

[207]*See, e.g.,* David Trafimow, *Editorial*, 36 BASIC & APPLIED SOC. PSYCHOL. (2014); David Trafimow & Michael Marks, *Editorial*, 37 BASIC & APPLIED SOC. PSYCHOL. (2015).

[208]Ronald L. Wasserstein & Nicole A. Lazar, *The ASA's Statement on p-Values: Context, Process, and Purpose*, 70 AM. STAT. (2016).

It does, however, go into great detail on proper use and interpretation of the *p*-value in the frequentist framework. Seeking relevance for statistical practice across scientific and policy-related disciplines, the statement emphasizes that "[w]hile the *p*-value can be a useful measure, it is commonly misused and misinterpreted."[209] The statement concludes with a call for good study design and scientific inference, closing: "No single index should substitute for scientific reasoning."[210]

B. Alternatives to the P-Value Paradigm

In light of these challenges to the *p*-value, many alternatives have been proposed. Remaining in the frequentist setting, a common call is for point estimates of effects and confidence intervals to be placed as prominently as *p*-values. Stuart Pocock's 1983 *Clinical Trials: A Practical Approach* encourages following significance tests with confidence limits (now more commonly called confidence intervals) "to estimate the magnitude of improvement of one treatment over another."[211] Frank Yates had previously endorsed this in 1951, lamenting that Fisher's works had "caused scientific research workers to pay undue attention to the results of tests of significance" and "too little to the estimates of the magnitude of the effects they are investigating."[212]

The confidence limits provide a range of estimates of the true effect under study such that, if the experiment is repeated under identical conditions many times, the given percentage of such confidence intervals will include the true parameter. As Pocock notes, "[i]t is standard practice to use 95% confidence limits."[213] This practice is as arbitrary, and largely derived from, the practice of using an alpha level of 0.05 in significance tests. Specifically, the link between confidence intervals and significance tests is clear: if the 95 percent confidence interval includes the value suggested by the null hypothesis, then the equivalent test will not be statistically significant at the 0.05 level. This feature of confidence intervals is sometimes, though not always, noted by its proponents. The use of the confidence interval brings the magnitude of effects and clinical relevance to the fore, but it does not do away with the problems of frequentist inference and does not lead to a clear decision rule that is different from the significance testing regime. It is worth noting that confidence limits were put forth, although not in their modern nomenclature, by both Ronald Fisher and Jerzy Neyman in the 1930s.[214]

For those who more fundamentally object to the frequentist paradigm, Bayesian statistics offers an attractive alternative. While this article cannot provide a reasonable treatment of the fundamentals of Bayesian inference, there is a benefit for clinical trial use that is fairly easy to understand. In many forms of Bayesian inference, an investigator begins with a prior belief about the parameter, or estimator, that she wishes to estimate (say, the difference in survival probabilities of those on a new treatment compared to those on a placebo). This prior belief is "updated" with

---

[209] *Id.* at 131.

[210] *Id.* at 132.

[211] POCOCK, *supra* note 98, at 206.

[212] Yates, *supra* note 200, at 32. Fisher's student L. H. C. Tippett had also warned that the "statistical significance gives no information as to the magnitude or practical importance of any difference." TIPPETT, *supra* note 85, at 51.

[213] POCOCK, *supra* note 98, at 208.

[214] *See* DAVID & EDWARDS, *supra* note 100, at 189.

information contained in the collected data through the use of the likelihood of that data appearing given the possible parameter values. Incorporating (formally, conditioning on) this data then gives an updated, or posterior, probability distribution for the parameter. This distribution can be described in many ways, with point estimates for the parameter given by the mean, median, or mode of the distribution and measures of uncertainty given by the standard deviation of the distribution or an interval in which the parameter has a certain probability of falling.[215] A Bayes Factor can also be calculated, which is the ratio of the posterior odds of one hypothesis and the prior odds of a competing hypothesis (usually the alternative hypothesis compared to the null hypothesis).[216] This quantity has been referred to as a measure of the strength of the evidence contained in the data and is sometimes put forth as an alternative to the *p*-value.[217]

The Bayesian alternative provides intuitive results. The parameter does have a 95 percent probability of being within the 95 percent credible interval, unlike a frequentist 95 percent confidence interval, in the strictest sense. Since *p*-values are often mistakenly interpreted to indicate the probability of the parameter having some value given the data, the use of Bayesian results can be easier to explain.[218] Additionally, the use of a prior distribution for the parameter allows the investigator to incorporate past information into the model. This has led to some criticism of Bayesian inference, however, as it has been accused of being open to subjectivity on the part of the investigator in selecting the prior belief.[219] Because of this, some Bayesian proponents have proposed the use of non-informative priors, although this often leads to numerical results identical to those from frequentist inference. The use of sensitivity analyses, common for various reasons in statistical analysis of biomedical data, is often proposed to determine the influence of a prior.[220]

C. FDA's Response to the Alternatives

FDA has encouraged the use of confidence intervals to represent trial results, but primarily as a supplement to testing procedures. In 1988, the Guideline for the Format and Content of the Clinical and Statistical Sections of an Application made some references to (usually 95 percent) confidence intervals as supplements to point estimates and its example tables include some confidence intervals around estimates. These do not take the place of *p*-values in reporting trial results, however, but

---

[215]The debate between frequentist probability and Bayesian inference has been going on for centuries, and the author has no designs on tackling the subject here; various books and articles explain both the fundamentals of Bayesian inference and the benefits and drawbacks of the Bayesian and frequentist frameworks. *See, e.g.,* GELMAN ET AL., *supra* note 33, at 3–32 (covering in a fairly accessible manner the statistical basis for Bayesian inference); Ronald A. Fisher, *Inverse Probability*, 26 MATHEMATICAL PROC. CAMBRIDGE PHIL. SOC'Y (1930) (remarking on the history of the Bayesian approach and critiquing that framework).

[216]GELMAN ET AL., *supra* note 33, at 184–86.

[217]Steven N. Goodman, *Toward Evidence-Based Medical Statistics. 2: The Bayes Factor*, 130 ANNALS INTERNAL MED. (1999).

[218]Duminda Wijeysundera et al., *Bayesian Statistical Inference Enhances the Interpretation of Contemporary Randomized Controlled Trials*, 62 J. CLINICAL EPIDEMIOLOGY (2009).

[219]*See, e.g.,* Goodman, *supra* note 217.

[220]*Id*.

supplement them.[221] The agency's 1998 Guidance for Industry on Statistical Principles for Clinical Trials gave fairly strong support to confidence intervals as supplements to significance tests and point estimates. "Estimates of treatment effects," the guidance states, "should be accompanied by confidence intervals, whenever possible."[222] In the 2015 guidance, the agency explicitly stated that sponsors should include estimated effect sizes, confidence intervals, and *p*-values in submissions: "A presentation of *p*-values alone would not be adequate."[223] Very narrow confidence intervals are mentioned as part of the statistical evidence that can lead to approval with one trial in the May 1998 guidance document as well.[224]

Despite these supportive statements around confidence intervals, they have not supplanted *p*-values in significance determinations. This may be due in part to the fact that confidence intervals do not lend themselves to binary decision-making (except insofar as they are equivalent to hypothesis tests) and so are more appropriate for the building of evidence that occurs in medical literature, which generally "requires no firm decision."[225] In addition, confidence intervals depend heavily on the size of an effect, which is largely why many practitioners prefer them. But FDA has generally interpreted its efficacy requirement to show "substantial evidence" that the drug has its purported effect, regardless of the magnitude of that effect, and distinctions between statistical significance and clinical significance are often blurred to the point of nonexistence.[226] Because of this, a *p*-value may be more appropriate for the evidentiary requirement than a confidence interval. Additionally, turning to 95 percent confidence intervals would not diminish the reliance on the 0.05 figure itself.

The Bayesian framework has generated considerable interest in the clinical trials field, especially over the last twenty years. Bayesian analyses of efficacy have been accepted for approval of drugs already (e.g., Pravigard Pac for prevention of myocardial infarction) and some trials are underway with prospective Bayesian designs.[227] In many ways, the medical devices field has led the way on the use of Bayesian trials. Certain classes of medical devices are subject to premarket approval by FDA, but the standard for proving effectiveness is not the same as for drugs and, in fact, "seems more flexible than the 'substantial evidence' standard applicable to drugs."[228] Because of this, there has been more room for alternative statistical techniques, including Bayesian methods to gain ground.

In the drug sphere, the 1998 Statistical Principles for Clinical Trials noted that it focused on frequentist methods but that "[t]his should not be taken to imply that other approaches are not appropriate; the use of Bayesian . . . and other approaches may be considered when the reasons for their use are clear and when the resulting conclusions

---

[221]U.S. FOOD & DRUG ADMIN., *supra* note 153, at 39, 66, 68.

[222]U.S. FOOD & DRUG ADMIN., *supra* note 157, at 32.

[223]U.S. FOOD & DRUG ADMIN., *supra* note 159, at 8.

[224]U.S. FOOD & DRUG ADMIN., *supra* note 171, at 15.

[225]Jonathan A. C. Sterne & George D. Smith, *Sifting the Evidence—What's Wrong with Significance Tests?* 322 BRIT. MED. J. 226, 229 (2001).

[226]Darrow, *supra* note 22, at 2125–26.

[227]*See, e.g.*, Donald A. Berry, *Bayesian Clinical Trials*, 5 NATURE REVS. DRUG DISCOVERY 27, 27 (2006); Donald A. Berry, *Adaptive Clinical Trials: The Promise and the Caution*, 29 J. CLINICAL ONCOLOGY 606, 606–07 (2011) [hereinafter Berry, *Adaptive*].

[228]HUTT ET AL., *supra* note 5, at 1236.

are sufficiently robust."[229] In 2004, FDA and Johns Hopkins University jointly held a conference entitled "Can Bayesian Approaches to Studying New Treatments Improve Regulatory Decision Making?", which covered drugs as well as devices. The August 2005 issue of the journal *Clinical Trials* covered this workshop, publishing many of the talks given there.[230] In her remarks at the workshop, then-Acting Deputy Commissioner for Operations at FDA, Janet Woodcock, encouraged participants to "push forward in the Bayesian area."[231] Robert Temple, then the Director of the Office of Medical Policy at CDER, took a somewhat more cautious tone. While expounding on the ways in which prior information was incorporated into FDA approval processes, including by lowering the *p*-value standards that might be required for a single trial, he noted that explicit Bayesian proposals were still very rare for drug trials. And he warned against generalizing from the experience of device approvals, noting that device manufacturers may be "much more prepared to make assumptions about what to expect. It is therefore not really clear that the CDRH (Center for Devices and Radiation Health) studies and drug studies are exactly the same in that sense."[232] Despite promise in the device field, then, and a workshop and an issue of a prominent journal devoted to explicating Bayesian trial approaches, the prospects were not necessarily bright for a wholesale renovation of the drug approval architecture.

Since then, there has been some uptake of Bayesian designs, but it remains limited. The biggest growth has been in the use of Bayesian methods in adaptive clinical trials. As adaptive designs have grown to get the most information out of a limited number of trial participants, the Bayesian analysis methods that conform well to updating with additional data have been frequently (but not always) paired with these designs.[233] Draft guidance issued in February 2010 entitled "Adaptive Design Clinical Trials for Drugs and Biologics" gave strong indications that FDA was open to adaptive designs and expected to work with sponsors to craft design and analysis plans that could lead to approval. But while it acknowledged the value of Bayesian analysis methods in these designs, it fell back to the old standard with regards to alpha levels: "In general, the study design should be planned in a frequentist framework to control the overall study Type I error rate."[234] While Bayesian methods were acknowledged, FDA was not ready to part with the significance testing framework that had defined drug approvals for decades. To date, this draft guidance has not been finalized; finalized guidance on adaptive designs for device trials issued in July 2016 was slightly more favorable to Bayesian methods but also recommended controlled Type I error rate.[235]

While statisticians, biomedical investigators, and journal authors in related fields have debated the proper role of *p*-values, significance testing, and the entire frequentist

---

[229]U.S. FOOD & DRUG ADMIN., *supra* note 157, at 4.

[230]Norris E. Alderson, *Introduction*, 2 CLINICAL TRIALS 271, 271 (2005).

[231]Janet Woodcock, *FDA Introductory Comments: Clinical Studies Design and Evaluation Issues*, 2 CLINICAL TRIALS 273, 275 (2005).

[232]Temple, *supra* note 108, at 281.

[233]Berry, *Adaptive*, *supra* note 227, at 606. *See also* Shein-Chung Chow & Mark Chang, *Adaptive Design Methods in Clinical Trials—A Review*, 3 ORPHANET J. RARE DISEASES 1, 1–2 (2008).

[234]U.S. FOOD & DRUG ADMIN., GUIDANCE FOR INDUSTRY: ADAPTIVE DESIGN CLINICAL TRIALS FOR DRUGS AND BIOLOGICS 34 (2010).

[235]U.S. FOOD & DRUG ADMIN., ADAPTIVE DESIGNS FOR MEDICAL DEVICE CLINICAL STUDIES: GUIDANCE FOR INDUSTRY AND FOOD AND DRUG ADMINISTRATION STAFF 14 (2016).

framework, these techniques have largely remained the law of the land for FDA. Several new approaches have been used to some degree, but none have led to a substantial reduction in the reliance on *p*-values and Type I error control. This has not stopped academics and trialists from proposing further refinements, however. Some, like the "split-sample analysis" process proposed by Mark van der Laan and co-authors, transform how error is controlled but remain rooted in the significance testing framework and appeal to overall 0.05 alpha levels.[236] This process uses some trial data as an exploratory set to identify subgroups on which the drug may be safe and efficacious. The remaining data are used to confirm safety and efficacy in these subgroups, with a higher-than-usual statistical significance standard applied to control the overall Type I error rate.[237] Others, like the Bayesian Decision Analysis framework proposed by Leah Isakov and co-authors, explicitly break from past notions of error control and focus on some external standard for appropriate decision-making mechanisms.[238] This proposal assigns a cost to making an incorrect decision in the drug approval process based on the burden and severity of the disease in question and the safety and efficacy profile of the drug. A Bayesian framework is then used to determine the appropriate drug-specific threshold for the efficacy significance level for approval.[239] Given the reluctance to break from established (and well-understood) standards, however, it seems unlikely that any of these will see considerable use in the near future.

D. Patient Advocacy and Challenges to the FDA Regulatory Paradigm

In addition to the technical and statistical challenges, patients and their advocates, as well as industry-related voices, have challenged FDA's statistics-based approach to drug regulation. These objections reflect the tension between FDA's mandate to ensure the safety and efficacy of drugs sold in the United States and the goal of patients, their advocates, and the companies manufacturing and selling pharmaceuticals to ensure that therapies move from the lab to the consumer as quickly as possible and that scientists and companies are incentivized to generate new therapies.[240]

This opposition has a long history, and often depends on very specific circumstances or specific drugs. At the time of the passage of the Kefauver-Harris Amendments, the medical profession opposed the new powers bestowed upon FDA, the drug industry "acquiesced reluctantly," and "organized consumer groups heartily endorsed it."[241] As soon as FDA began putting regulations into effect, however, patients and doctors responded. When the Drug Efficacy Study recommended the

---

[236]Mark Van der Laan et al., *Improving the FDA Approval Process* at 5 (John M. Olin Program in Law and Economics Working Paper No. 580, 2011), http://chicagounbound.uchicago.edu/law_and_economics/256/ [https://perma.cc/5G9H-U2PS].

[237]*Id.*

[238]Vahid Montazerhodjat & Andrew W. Lo, *Is the FDA Too Conservative or Too Aggressive?: A Bayesian Decision Analysis of Clinical Trial Design* 1, 1–2 (NBER Working Paper No. 21499, 2015), http://www.nber.org/papers/w21499 [https://perma.cc/8R38-YZZ8].

[239]*Id*. at 14.

[240]*See, e.g.,* Kulynych, *supra* note 22, at 127–28 (detailing the debate over this tension during debate over the 1997 FDA Modernization Act).

[241]Irving H. Jurow, *The Effect on the Pharmaceutical Industry of the "Effectiveness" Provisions of the 1962 Drug Amendments*, 19 FOOD DRUG COSM. L.J. 110, 112 (1964).

removal of bioflavonoids from the market, consumers and their doctors began an intensive, though ultimately unsuccessful, lobbying effort. Foreshadowing many future arguments, one patient wrote to his senator about his concern at the regulators "countermand[ing] instructions of my personal physician of almost twenty years."[242]

This calculus has shifted repeatedly over time, but there are now very strong patient advocacy groups that oppose FDA decisions limiting access to new medicines. Health advocacy organizations exist for nearly every disease and medical condition, and can vary drastically in size and scope. Their main activities include the promotion of medical research, conducting disease awareness campaigns, and advocating "for policies that they believe are in their members' best interests."[243] Advocacy around clinical trials and FDA regulation increased drastically during the AIDS crisis, when patients and their supporters aggressively challenged FDA to speed up drug approvals and expand the rights of patients to try experimental therapies.[244] As activities around drug access increased, some advocacy organizations began to partner with pharmaceutical companies, including by accepting funding from the corporations.[245] Other organizations advocated for agendas similar to those of the companies without forming explicit partnerships.[246] With aligned interests in speeding the time to market of new therapies, patient advocates and industry representatives both took issue with some of the statistical approaches undertaken by FDA, including the use of significance tests.

In 1987, FDA faced one of the first of a new class of large-molecule biological products, drugs derived from—or synthesized to replicate—complex natural substances whose structure cannot be readily determined, when biotechnology company Genentech submitted an application for approval of tissue plasminogen activator (TPA). In May, FDA refused to approve the TPA submission, instead requesting more data on drug efficacy and safety.[247] The advisory panel recommended this step in large part because of the statistical reviewer's determination that Genentech's studies failed to show a "measurable, beneficial effect" of the drug.[248] This decision was met with derision from the scientific and lay media.[249] The *Wall Street Journal* editorial page led the charge with the most emotional attacks on FDA's decision and FDA official Robert Temple himself. "Medical research has allowed statistics to become the supreme judge of its inventions," wrote the editors, who went on to ask, "Are American doctors going to let people die to satisfy the bureau of drugs' chi-square studies?"[250] Although FDA approved the drug later that year when

---

[242] CARPENTER, *supra* note 112, at 350.

[243] Sheila M. Rothman et al., *Health Advocacy Organizations and the Pharmaceutical Industry: An Analysis of Disclosure Practices*, 101 AM. J. PUB. HEALTH 602, 602 (2011).

[244] *Id.* at 603.

[245] *See, e.g., id.* at 606–07.

[246] *See, e.g.,* CARPENTER, *supra* note 112, at 393–461.

[247] CARPENTER, *supra* note 112, at 2–5.

[248] Peter R. Kowey et al., *The TPA Controversy and the Drug Approval Process: The View of the Cardiovascular and Renal Drugs Advisory Committee*, 260 J. AM. MED. ASS'N 2250, 2251 (1988).

[249] *See, e.g.,* Daniel E. Koshland, Jr., *TPA and PDQ*, 237 SCIENCE 341, 341 (1987). *See also* CARPENTER, *supra* note 112, at 4 (describing several other critical articles, editorials, and statements by scientists).

[250] Editorial, *Human Sacrifice*, WALL ST. J., Jun. 2, 1987, at 30.

presented with further study results,[251] the TPA question demonstrated the intense controversies that accompany any FDA drug rejection on efficacy grounds.

Even after the 1997 FDA Modernization Act helped speed up and expand access to therapies,[252] the debate over the proper role of FDA as gatekeeper to therapies has not abated. Controversies over drug approvals (or, more commonly, non-approvals) have continued, often led by patient advocacy groups. In particular, these advocacy groups raise concerns about Type II errors in drug decisions, FDA failing to approve a drug when it does in fact have an effect. Advocates, especially in the context of severe diseases with limited treatment options, point to unnecessary disease burdens and death because of these Type II errors.[253] As discussed *supra* section III.A, reducing Type II errors in statistical analyses would necessarily mean increasing the risk of Type I errors by raising the alpha level and approving more drugs that may be ineffective.

The strict adherence to a 0.05 significance level may be unpalatable to advocates of particular treatments because of the risk of Type II errors. On the other hand, given its statutory mandate to ensure that drugs are approved only with "substantial evidence that the drug will have the effect it purports or is represented to have," FDA must create some standard by which to assess the statistical evidence provided by clinical trials. With the biomedical community having come to accept hypothesis testing in the 1940s and 1950s, and Fisher's 0.05 level, though arbitrary, becoming commonly used, any other level would be difficult to defend. Additionally, drug sponsors generally crave consistency and some foreknowledge that their clinical trial plan will lead to a positive result if the statistics meet the agreed-upon level, leading them to take FDA guidance very seriously and interact frequently with FDA officials.[254] To change the standard now would disrupt not only FDA's processes for reviewing clinical trial data, but also sponsors' processes for planning and analyzing trials, and public confidence in FDA's past and future drug decisions. Any standard that did not rely on a specific significance level threshold would be open to challenges of subjectivity or inconsistent application, creating a much more difficult approval scheme for FDA to defend and justify statutorily.

CONCLUSION

Over fifty years after the Kefauver-Harris Amendments created the drug efficacy review regime and nearly twenty years after the last major statutory change came with FDAMA, Section 505 was amended again. The 21st Century Cures Act, (Cures Act), an omnibus biomedical research bill passed in December 2016, introduced a number of novel ideas around the "substantial evidence" standard. The largest is the new Section 505(f), entitled "Real World Evidence," which directs the Secretary of Health and Human Services to "evaluate the potential use of real world evidence" in drug approval processes and post-approval surveillance.[255] Real world evidence is then

[251] Kowey et al., *supra* note 248, at 2252.

[252] Kulynych, *supra* note 22, at 127.

[253] Daniel P. Carpenter, *The Political Economy of FDA Drug Review: Processing, Politics, and Lessons for Policy*, 23 HEALTH AFFS. 52, 57–58 (2004).

[254] Temple, *supra* note 18, at 1646–47.

[255] Pub. L. No. 114-255, 34 H.R. 64 (2016) (codified at 21 U.S.C. §355(f)).

defined as "data regarding the usage, or the potential benefits or risks, of a drug derived from sources other than randomized clinical trials."[256] A few paragraphs later, however, the bill clarifies that it "shall not be construed to alter . . . the standards of evidence under . . . section 505, including the substantial evidence standard in such subsection (d)."[257]

As the bill was nearing passage, then-FDA Commissioner Robert Califf and colleagues wrote in the *New England Journal of Medicine* about the promises and perils of "real world evidence." They urged "caution" and tempering of "expectations of 'quick wins,'" stating that "[r]eal-world research and the concepts of a planned intervention and randomization are entirely compatible."[258] With no regulations or guidance documents yet promulgated expounding on the "real world evidence" mandate, it is unclear exactly what effect the new legislation will have on the drug approval standards. But some health policy researchers have warned that the legislation may "encourage use of less rigorous data to meet standards for approval."[259] Others further suggest that new standards are already starting to influence FDA approvals, citing Sarepta Therapeutics' Duchenne's Muscular Dystrophy drug as "the accelerated approval of a drug with inadequate clinical trials and weak efficacy data."[260]

In 2017, Portola Therapeutics decided to test the willingness of the new presidential administration and its FDA leadership to be flexible with statistical cutoffs. The drug in question, betrixaban, showed promise in early-stage trials in preventing blood clots after an illness. In its pivotal phase 3 trial, however, betrixaban failed to meet the pre-specified 0.05 alpha level for the primary endpoint, showing a *p*-value of 0.054.[261] Portola pointed FDA towards an alternative interpretation of the results, however, and won approval in late June.[262]

While the Portola case may indicate some changing standards at FDA, the further effects of the Cures Act and new FDA Commissioner Scott Gottlieb remain to be seen. But the use of *p*-values, significance testing, and the 0.05 alpha level at FDA have fifty years of history. They have survived in-fighting among the pioneers of the statistical methods, accusations of arbitrary cutoffs, a push for Bayesian statistics, a rejection of

---

[256] *Id.*

[257] *Id.*

[258] Rachel E. Sherman et al., *Real-World Evidence—What Is It and What Can It Tell Us?*, 375 NEW ENGL. J. MED. 2293, 2296 (2016).

[259] Aaron S. Kesselheim & Jerry Avorn, *New "21st Century Cures" Legislation: Speed and Ease vs Science*, 317 J. AM. MED. ASS'N 581, 582 (2017).

[260] Amitabh Chandra & Rachel E. Sachs, *An FDA Commissioner for the 21st Century*, 376 NEW ENGL. J. MED. e31(1), e31(2) (2017).

[261] Alexander T. Cohen et al., *Extended Thromboprophylaxis with Betrixaban in Acutely Ill Medical Patients*, 375 NEW ENGL. J. MED. 534, 534 (2017).

[262] Letter from Dr. Richard Pazdur, U.S. FOOD & DRUG ADMIN., to Janice Castillo, Portola Pharmaceuticals, Inc. (Jun. 23, 2017), https://www.accessdata.fda.gov/drugsatfda_docs/appletter/2017/208383Orig1s000ltr.pdf [https://perma.cc/J6B6-RMCR]. *See also* Adam Feuerstein, *Will Trump's FDA Relax Standards for Drug Approval? We'll Get a Clue Soon with Decision on Portola's Anticoagulant*, STAT NEWS (Jun. 21, 2017), https://www.statnews.com/2017/06/21/portola-drug-fda-decision/; Damian Garde & Adam Feuerstein, *Surprise Approval for Portola's Blood Thinner has Biotech Eyeing a New Dawn at the FDA*, STAT NEWS (Jun. 23, 2017), https://www.statnews.com/2017/06/23/portola-fda-approval/ (describing the evidence accumulated by Portola, the controversies surrounding the approval process, and potential future ramifications of the approval decision).

p-values in some disciplines and reassessment of their use in others, patient, physician, and pharmaceutical company challenges, and countless administrations and FDA officials. The *p*-value itself has been used to assess evidence for three hundred years and is intimately associated with the rise of the modern randomized clinical trial; the 0.05 significance level came with the *p*-value into the biomedical world from Fisher's works on the subject. Adapted and re-assessed throughout the decades to meet the changing needs of FDA and respond to statistical and biomedical advances, this framework remains a cornerstone of U.S. pharmaceutical policy and looks poised to remain so for the foreseeable future.

"Almost every phase of the practice of medicine necessitates at least the rudimentary application of statistical ideas."[263] As true as that statement by sociologist and pioneering National Cancer Institute epidemiologist Harold Dorn was in 1955, on the eve of FDA's turn to statistical evidence, it is even more true today.

[263]Harold F. Dorn, *Some Applications of Biometry in the Collection and Evaluation of Medical Data*, 1(6) J. CHRONIC DISEASES 638, 638 (1955).