Spring 2018

The Weight of a Term: "Substantial Evidence" and Buried Data

Stewart Justman
University of Montana - Missoula, stewart.justman@umontana.edu

The Weight of a Term

"substantial evidence" and buried data

Stewart Justman

ABSTRACT When Congress amended the Food, Drug, and Cosmetic Act in 1962 to ensure the efficacy of drugs before they reach the market, it imposed a standard of evidence distinctly weaker than the reasonable one of preponderance. The difference is material. Some drugs now on the market may or may not have a preponderance of trial evidence of efficacy. An obstacle to a finding of efficacy is the well-known nemesis of drug makers with a definite interest in favorable trials: the placebo effect. Trials where the placebo effect runs high are vulnerable to negative findings, and trials with such findings all too often find their way to burial sites like regulatory archives. Publication bias—the preferential reporting of positive findings—has been abetted by a regulatory standard that does not require a preponderance of evidence and discounts negative trials, provided only that two trials show positive results.

It's common knowledge that published medical findings tell only part of the story, the other part consisting of unpublished records of negative trials, sometimes of the same agents. No one, of course, knows precisely how much lies below the waterline, but informed estimates suggest that as many as half of all trials undertaken remain unreported (Ioannidis 2017)—a percentage sufficient to cast a shadow over the entire corpus of published medical research. The effect is incal-

Emeritus Professor, College of Humanities and Sciences, University of Montana, 1703 Maurice Avenue, Missoula, MT 59801.

Email: stewart.justman@umontana.edu.

The author thanks Susanne Caro, Suzanne White Junod, and Michael Mayer for their assistance.

Perspectives in Biology and Medicine, volume 61, number 2 (spring 2018): 201–214.

culable. That negative findings of trials of cancer chemotherapies would be withheld from publication seems inconceivable, but a paper documented that practice, and its slanting effect on the published literature, some 30 years ago (Simes 1986). Severe enough to distort the medical literature and stubborn enough to persist for decades after exposure, the problem of the selective publication of positive findings—publication bias—must have deep roots. The favored explanation of publication bias is the heavy influence of corporate funders of clinical trials that use the medical literature to promote their products. Without foreclosing that explanation, I suggest a different, but also deeply entrenched, contributing cause: the rules governing drug approval enacted in 1962. Investigators who put negative findings into cold storage do no more than the FDA itself in many cases.

The Food, Drug, and Cosmetic Act of 1938 tasked the FDA with ensuring the safety of new drugs before they reach the market. With the revision of the Act in 1962 in the Kefauver-Harris Drug Amendments, the FDA was charged with ensuring the efficacy of new drugs as well. During the pharmaceutical boom of the 1950s, hundreds of new drugs or combinations poured into the marketplace each year, many of doubtful efficacy even though highly advertised to doctors. In what now looks like a commercial form of suppressed evidence, such ads often omitted side effects and even quoted endorsements of a drug with the original reservations left out (Lasagna 1962; McFadyen 1973). In an effort to check misleading advertising as well as exorbitant prices and other abuses, Sen. Estes Kefauver launched hearings into the drug industry in December 1959 (Greene and Podolsky 2012). However, not until 1962, when news of the thalidomide catastrophe broke, was congressional action assured. As the world learned of pregnant women who took thalidomide for morning sickness only to give birth to severely deformed infants, reform that once seemed unlikely became inevitable.

Compared with such red-hot issues as price and patent reform and with such an anguishing crisis as the thalidomide disaster, the question of drug efficacy might seem academic. Nevertheless, it commanded legislative and executive attention. In testimony before a House committee on June 20, 1962—just weeks before *thalidomide* became a household word—Abraham Ribicoff, Secretary of the Department of Health, Education, and Welfare (parent of the FDA), and George Larrick, FDA Commissioner, both maintained that drugs should be *proven* effective before being cleared for the market. What degree or proportion of evidence is required to meet this burden of proof? Presumably most of the evidence. The Kennedy administration accordingly drafted a bill with a key clause calling for approval of new drugs if and only if their efficacy and safety were demonstrated by a preponderance of the evidence, the customary administrative-law standard (Carpenter 2010). This reasonable provision was not to the liking of the American Medical Association (AMA) and the Pharmaceutical Manufacturers Association (PMA), however, and in the course of the bargaining that fashioned a bill soon adopted unanimously by both houses of Congress, the pre-

ponderance standard gave way to "substantial evidence" in the form of "adequate and well-controlled investigations."[1] Lost in the glare of the latter phrase, which has come to be read as a kind of authority for the randomized clinical trial (RCT) and as the writ of the FDA as the trial's regulator, is the low level of evidence needed to show efficacy.

Politically speaking, the abandonment of the preponderance standard by the Kennedy administration was less dramatic than it seems, as the administration was cautious in its enthusiasm for drug reform, preferring to lead from behind and keeping its distance from Kefauver throughout his hearings. Practically speaking, the loss of preponderance left a void that the category of "substantial evidence" could not fill. Note that in the context of the Drug Amendments, "preponderance of evidence" and "substantial evidence" are not even parallel terms, the former referring to a proportion of evidence, the latter to the method by which evidence is gathered ("adequate and well-controlled investigations"). How much evidence constitutes "substantial evidence"? According to one present at the creation of the Drug Amendments, the term means "more than a scintilla of evidence but not necessarily a preponderance" (Lasagna 1989).[2] Even if those who originally argued for "substantial evidence" wouldn't really have placed it at the low end of this expansive scale, clearly a world of difference separates the two standards.

Two Standards

In sending the Drug Industry Act of 1962 to the floor of the Senate on August 21, 1962, Sen. James Eastland, chair of the Judiciary Committee, reported that "in the course of committee deliberations a distinction evolved . . . between two tests—the 'preponderant evidence' test and the 'substantial evidence' test," and that the committee adopted the second. Congress would follow suit. However, in rejecting preponderance in favor of "substantial evidence," the various framers of the Drug Amendments replaced an appropriate measure of evidence with a term that doesn't really measure anything. Possibly just because of the insubstantiality of "substantial evidence," the FDA has focused instead on the word *investigations*, which it interprets to mean more than one investigation.[3] Not the

[1]The statute language reads, "As used in this section and subsection, the term 'substantial evidence' means evidence 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 reasonably 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."

[2]Lasagna paraphrased the courts. See Laws v. Celebrezze, 368 F. 2d 640, Court of Appeals, 4th Circuit (1966): "Substantial evidence . . . consists of more than a mere scintilla of evidence but may be somewhat less than a preponderance."

[3]Warner-Lambert Co. v. Heckler, 787 F. 2d 147 (1986). Under the 1997 revision of the Food, Drug, and Cosmetic Act, the FDA was authorized to accept the findings of one positive trial under certain circumstances. See Anello and Junod 2013; Kesselheim and Darrow 2015.

important-looking word *substantial* but the letter *S* sets the level of evidence for drug approval. Operationally, "substantial evidence" translates into the positive findings of the minimum plural number of well-conducted trials—two—regardless of evidence to the contrary. Whereas a preponderance standard "would have required a majority of a company's studies to have demonstrated positive results" (Carpenter 2010, 258), an FDA-approved drug may or may not have a majority of positive trials under the two-trials requirement. While this requirement has been acclaimed as "revolutionary" (Anello and Junod 2013), revolutions rewrite memory, and the price at which the two-trials principle was purchased should not be forgotten.

Far from being demanding, "substantial evidence of efficacy" represents a comparatively weak standard—weak enough, in any case, for Pres. Kennedy to have objected in a letter to Sen. Eastland on August 3, 1962, that "this standard of proof is inadequate in terms of assuring that drugs that reach the market have been shown to be effective for the claims made for them." The key word, *substantial*, is a term of recognized laxity. As noted by the Court of Appeals, 4th Circuit, Congress "specifically discarded" the standard of preponderant evidence "for the milder term 'substantial.'"[4] Indeed, the courts regard "substantial evidence" as the lowest evidentiary standard (Darrow 2013), a status that makes its substitution for preponderance the equivalent of a vertical drop. The relative permissiveness of "substantial evidence" no doubt eased the passage of a law which, not long before, faced the determined opposition of the pharmaceutical industry and the AMA combined. And the effects of the lower standard continue to be felt. Recently an FDA official mused aloud on the possibility of a drug with two positive studies "out of 14 or 12," a result that could well occur by chance alone (Firth 2015). While that would be an unusual case, the efficacy language now in force silently implies the possibility of negative trials, much as the published medical literature silently implies an unpublished one. Though "companies cannot 'cherry-pick' the two best trials to submit to the FDA while withholding the others from its consideration" (Darrow 2013, 2095), such consideration may well result in the archiving of the unsuccessful trials, not the disapproval of the drug. How ironic that the two-trials rule, established on the basis of an inoperative adjective and a plural noun, acquired the same nickname as the institution of the RCT itself: the "gold standard" (Kulynych 1999).

Aside from their difference in clarity and rigor, the preponderance criterion takes failures into account while "substantial evidence" does not. Under the former, if there are too many failures, the drug lacks efficacy. Before the preponderance standard was bargained away, one of its adherents wanted to make sure that "if a hundred doctors tested [a] drug and only twenty considered it effective," the drug would be found ineffective (Harris 1964, 204). (One wonders what the tests consist of, though.) Even in less one-sided cases, the balance of the evidence

[4]Hynson, Westcott and Dunning, Inc. v. Richardson, 461 F.2d 215 (1972).

necessarily matters under a preponderance standard. The "substantial evidence" standard as put into practice writes off negative trials provided only that two trials show positive outcomes.[5] With the second such trial, it seems, a threshold is crossed and a result enters the prestigious category of a reproduced finding.

Originally it took nine trials of paroxetine to yield two positives (Fava et al. 2003). Under an efficacy standard that accepts a showing like this as substantial evidence in favor of the test drug, the negative data often ends up archived because it has no bearing on the approval of the drug. We have reason to believe, then, that an official standard of evaluation that renders negative trials moot has something to do with their relative exclusion from the medical literature. Some might say that burial in the files of a regulatory agency, or simply an office drawer, is a fitting fate for data that lacks evidentiary import and does not factor into official judgments of efficacy. That Congress foresaw such a loss of data even as it reacted to the thalidomide crisis seems improbable.

A Low Bar

Given that Kefauver himself, the Drug Amendments' prime mover, had taken the position that a drug should show substantial evidence of efficacy, that standard was only too available as a fallback position when the time came to enact a law.[6] And yet as legislative matters heated up in June 1962, Kefauver objected to a proposal by two members of the Senate Judiciary Committee to replace the requirement that a patentable drug show "greater therapeutic effect" with the requirement that it offer only a "substantial" improvement: "This wording seemed imprecise to Kefauver, and he was insistent that the original wording be preserved. . . . To Kefauver the point was crucial" (Harris 1964, 162). On the also crucial matter of drug efficacy—crucial enough to shape the official standard of drug evaluation from 1962 to this moment—the same imprecision prevailed. It is hard to avoid the conclusion that Congress adopted the "substantial evidence" standard precisely because of its laxity.

The creation of a system of drug approval leaving broad scope for mixed results appears to have been entirely deliberate. As an FDA historian has written,

> The legal language employed in the [Drug Amendments of 1962], which laid out the criteria that would be used in assessing efficacy in support of a new drug approval, was not particularly stringent. The law required that there be "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." Lawyers have concluded that Congress could have

[5]On the FDA's discretionary waiver of the two-trials requirement, see Kulynych 1999.

[6]In a colloquy on drug efficacy in a hearing of the Subcommittee on Antitrust and Monopoly, December 7–9, 1961, Kefauver said, "Substantial evidence is required in most Government procedures, and that is inherent in what we have in mind with the bills."

> established a more stringent drug approval process simply by using stronger legal terminology. The fact that terms such as "preponderance of evidence" or "evidence beyond a reasonable doubt" were not used indicates that Congress did not intend to set the bar for efficacious new drug approvals too high. New drugs did not have to be superior to other drugs on the market nor did "substantial evidence" mean evidence "so strong as to convince everyone." (Junod 2008)

Tellingly, no characterization of "substantial evidence" is offered except to say that it's weaker than a preponderance and doesn't necessarily win everyone over. As strangely vacant as the concept is, that is how the framers of the Drug Amendments understood it. In hearings on what became the Amendments, critics argued for an efficacy criterion whose operative definition was simply that it was not a preponderance of evidence. The following exchange in a hearing of the House Committee on Interstate and Foreign Commerce is characteristic:

> Mr. DINGELL. All right. Do you think that the efficacy should be supported just by substantial evidence or should it be supported by the preponderance of evidence?
> Dr. KEEFER. I think it should be supported by substantial evidence.

To its opponents, the preponderance standard was at once dictatorial and artificial. Such a delicate question as the efficacy of a drug is not to be decided by the tyranny of averages or the edicts of regulators and is best left to time and the free play of informed opinion. Indeed, critics one after another rejected the proposition that premarket trials alone can determine the efficacy of a drug. In the submitted opinion of the AMA, because "a drug's efficacy varies from patient to patient," only the prescribing physician can judge a drug's worth for any given patient. (This position appears to clash with one taken by the AMA a few years before, namely that "the average physician has neither the time nor the facilities to experiment with new drugs in order to determine their proper indications for use" [Wardell and Lasagna 1975, 14].) The true measure of a drug makes itself known not in a handful of clinical trials but in "the extensive use of the drug by large numbers of the medical profession over a long period of time." It was because such practical studies (if that is the word) could be expected to yield conflicting findings that many opposed a standard of efficacy that seemed to them unreasonably rigid. And if the rejection of preponderance opens the possibility of approval of a drug found effective only in a minority of studies, the framers did not shrink from this implication. Thus, in a 1962 report of the Senate Judiciary Committee we read:

> The term "substantial evidence" is used to require that therapeutic claims for new drugs be supported by reliable pharmacological and clinical studies. When

> a drug has been adequately tested by qualified experts and has been found to have the effect claimed for it, this claim should be permitted even though there may be preponderant evidence to the contrary based upon equally reliable studies.

To many who rejected the preponderance standard in favor of the looser criterion of "substantial evidence," it seems a trial or study was something conducted by the doctor in his office, without randomization, controls, or other procedural niceties. (Thalidomide itself had been given to American patients by their doctor as an "investigational" drug, as if handing out samples constituted a trial.) In the course of the Kefauver hearings, some witnesses complained of the evidentiary worthlessness of this sort of exercise. "I gave drug 'x' to patients 1 through 10, and they all did quite nicely on it" (Anello and Junod 2013, 2). The tradition of case studies survived the onslaught of the Kefauver hearings and the revision of FDA rules, and two decades later an acute sociologist renewed the criticisms of the genre. He also pointed out that such studies exhibit the telltale signs of publication bias: "Medical journals, along with those in other fields, tend to publish only reports of 'successful' interventions. One seldom reads of unsuccessful interventions, even though their frequency may be equal to, and probably greater than, those purported to be successful" (McKinlay 1981, 379). While RCTs represent an evidentiary advance over case studies, the abandonment of the preponderance standard in the course of the Drug Amendments' evolution into law invited a similar outcome. If a drug meets the two-trial requirement, negative trials are academic, after all; they become footnotes to history.

Over the course of the Kefauver hearings it emerged that pharmaceutical companies had much information about their drugs that was never disclosed to the FDA (Goodrich 1963). Over the decades to come, the FDA itself would amass trial data that was not disclosed to the public. Certainly this did not happen by Congressional design. Though the authors of the Drug Amendments evidently knew they were leaving latitude for unsuccessful trials, they did not realize that by depriving them of evidentiary significance they were creating the conditions for their burial.

The Challenge of the Placebo Effect

Opposition to the preponderance standard in 1962 seems to have arisen in part from a misunderstanding of that concept. Alleging that "if a drug company should be required to submit a preponderance of evidence as to the effectiveness of a new drug, it is highly doubtful that any drug which is effective in treating a limited percentage of patients would ever be marketed," critics argued that a drug shown to help 20% of patients with epilepsy would be disallowed because it failed to benefit the other 80%. In the world as we know it, a drug doesn't have to help a preponderance of patients in order to be found effective: "Countless drugs that have been shown in randomised controlled trials to be effective work

in only a minority of patients" (Christakis 2008, 1025). The question before most trials is not whether a drug helps more patients than not but whether it helps more patients than placebo; and if enough so benefit, the trial succeeds (provided of course that the drug doesn't do more harm than good). Not that success comes easily. While some hold that "better than placebo" simply means "better than nothing" (Avorn 2005, 970)—which would be the lowest standard imaginable—the power of the placebo has undone many a drug in preliminary testing and poses a nemesis for the companies that make them (Potter, Mallinckrodt, and Detke 2014). A recent review of published trials of drugs for neuropathic pain, for example, found a rising placebo response (linked to longer trials offering "richer social support") and a correlative dwindling of the drug-placebo difference (Tuttle et al. 2015). How confounding the placebo effect must be if "a dose of enthusiastic iatrotherapy [that is, comfort and reassurance] given in conjunction with an ineffectual drug will usually make patients feel much better than a moderately effective drug delivered with little or no iatrotherapy" (Feinstein 1971, 552). Inasmuch as trials that fail to establish the superiority of drug to placebo often fail to appear in the medical literature, the power of the placebo is highly implicated in the problem of publication bias.

The placebo effect bedevils in particular trials of psychoactive drugs, and for this reason "when a pharmaceutical company seeks FDA approval for a new drug treatment for psychiatric disorders, it is quite common for the NDA [New Drug Application] submission to be filled with trials where superiority over placebo was not demonstrated" (Fava et al. 2003, 116). That the published literature is filled with trials where superiority over placebo *is* demonstrated implies that much trial data has disappeared from public view. Perhaps the best-known excavation of data interred in the files of the FDA was the disclosure by Irving Kirsch in 2010 that more than half of trials of antidepressants reported to the FDA found no superiority of drug to placebo. As newsworthy as Kirsch's report was, the uneven showing of antidepressants was already a matter of record. In 2001 a meta-analysis of published and unpublished trials of tricyclic antidepressants against placebo found that the drugs failed to outperform placebo in 22 of 32 studies (Storosum et al. 2001): a preponderance of negative evidence. In 2002, a review of 15 years of data obtained from the FDA under the Freedom of Information Act found that antidepressants tested superior to placebo in 48% of trials (Khan, Khan, and Brown 2002). A 2008 review of published and unpublished studies of 12 antidepressants involving over 12,000 patients found that among those studies registered with the FDA, "whether and how the studies were published were associated with the study outcome," with the published literature suggesting that virtually all trials were positive, but FDA reviews showing a positive rate of 51% (Turner et al. 2008, 252).

The better to track the implications of decisions about standards of evidence made half a century ago, I have confined discussion to the question of efficacy,

putting safety to one side. Yet judgments of efficacy and harm intertwine, as in a recent notorious case of suppressed findings brought to light. A reanalysis of the data from SmithKline Beecham's Study 329, some of it contained in 77,000 pages of previously unpublished documents, revealed not only that paroxetine failed to outperform placebo, but that it increased the risk of "suicidal ideation and behaviour" in the study's adolescent subjects (Le Noury et al. 2015). (A review in *The Lancet* in 2005 concluded that more than half of all studies of antidepressants in children and adolescents failed to show superiority to placebo, with much data from negative trials "available only as posters from scientific meetings or . . . not yet available" [Ryan 2005, 938].) Given that paroxetine failed to outperform placebo in seven of nine trials in its original New Drug Application (as noted above), the negative finding of Study 329 is not in itself shocking, although it would seem to strain medical calculation to raise the risk of a suicide attempt in order to obtain the benefits of a placebo.

Paroxetine is also among the selective serotonin reuptake inhibitors (SSRIs) tested as a treatment for PTSD, with mixed results. An interesting report of an unsuccessful trial of the kindred drug fluoxetine for PTSD found that the placebo response rate ran "substantially higher than in a previously published fluoxetine trial of posttraumatic stress disorder" [*sic*]. The authors also note that in recent years several studies of SSRIs for PTSD "have been published with ambiguous results, negative or failed trials, and some studies with negative results remain unpublished" (Martenyi, Brow, and Caldwell 2007). As in the case of other studies of placebo-responsive conditions, the impression is of a continuum running from successful trials to equivocal ones to failures, with publication at one end and nonpublication at the other.

Fluoxetine in turn has also been studied as a treatment for premenstrual dysphoric disorder (PMDD). In the file reviewed by the FDA prior to the approval of fluoxetine (Sarafem) as a treatment for PMDD were (1) one trial in which the daily record of severity of problems declined 8% more for subjects on 20 mg of drug than on placebo; (2) one trial in which 2% more patients showed marked improvement on 20 mg of drug than on placebo; and (3) one trial in which 20 mg fluoxetine wasn't superior to placebo at all. (On the label from which I have taken this information, the FDA recommends a 20 mg dosage of Sarafem.) The latter trial is summarized by the FDA as follows:

> In another continuous dosing, double-blind parallel group study, patients with LLPDD [late luteal phase dysphoric disorder] (N=42) were treated daily with fluoxetine 20mg/day, bupropion 300mg/day, or placebo for 2 months. Neither fluoxetine nor bupropion was shown to be superior to placebo on the primary endpoint, i.e., response rate (defined as a rating of 1 [very much improved] or 2 [much improved]), possibly due to sample size.[7]

[7]See http://www.accessdata.fda.gov/drugsatfda_docs/label/2009/021860s005lbl.pdf.

Unlike the accompanying summaries of the positive trials (one of which involved 19 subjects), this skeletal précis gives no statistical breakdown. One suspects the study in question lies unpublished.

Similarly, in trials of the weight-loss drug Contrave (naltrexone and bupropion), subjects shed 4%, 3%, and 2% more weight with the drug than placebo, though treatment also included behavioral counseling, with those undergoing the most intensive counseling showing the best results.[8] The drug label mentions the COR-II trial but, in contrast to the three nominally positive studies, does not report its results. Presumably this is another case where a high placebo response not only overwhelmed the drug but entombed the trial's findings.

Let a final example of archived evidence serve for many. A 1999 review of trials of treatments of irritable bowel syndrome (a condition known to be receptive to the placebo effect) observes that "most therapies have so far been only marginally better than placebo" and that "many studies in the literature used far too few patients to show an effect unless this was extremely large." Given these shortcomings, the author concludes that a number of positive findings "were the result of chance or else many negative studies were never published." Here too, nonpublication of trial findings seems to be associated with a placebo effect high enough to rival the effect of a drug (Spiller 1999; see also Darrow 2013). Once again we may note that a treatment tenuously superior to placebo may or may not accrue a preponderance of positive trials even though it meets the criteria for approval.

At the time of the Drug Amendments of 1962, an important line of argument, advanced notably by Henry Beecher in his seminal paper "The Powerful Placebo" (1955), held that nothing but rigorous trials could determine whether a drug was more effective than placebo. While Beecher emphasized the necessity of randomization, double-blinding, and statistical validation, as well as the use of placebos in what came to be called "well-controlled investigations," he silently assumed another element of the trial regime: the due reporting of findings. If, as he says, "Many a drug has been extolled on the basis of clinical impression when the only power it had was that of a placebo," how could such an error be corrected unless the outcome of trials showing the drug's efficacy as tantamount to placebo were reported to the world? Concealment of unsuccessful trials can also mask the inefficacy of a drug that may be superior to placebo, but not impressively so. Ondansetron has been widely used as anti-emetic even though a systematic review concluded that it "does not prevent PONV [post-operative nausea and vomiting] very well," adding that "any publication bias would be expected to show that that ondansetron prophylaxis works even less well" (Tramer et al. 1997).

In the same year as the Drug Amendments, Beecher's sometime-collaborator Louis Lasagna (1962) defended the importance of experiments that prove un-

[8]See http://www.accessdata.fda.gov/drugsatfda_docs/label/2014/200063s000lbl.pdf.

successful. After all, if we know the outcome of a study beforehand, why run it? Two decades later the point had still not sunk in, and a proponent of evidence-based medicine contended that the term "negative trial" constitutes a kind of prejudicial misnomer, blocking the recognition that "all trials that have been well conceived and well conducted—whatever their results—represent positive contributions to knowledge" (Chalmers 1985, 1002). So loaded with misleading connotations of failure was the expression "negative trial" that the author proposed in the spirit of Swift to ban it. Is it too much to suppose that a regulatory standard that rewards positive findings and moots negative ones has helped make this term prejudicial?

Disregarded Data

At a critical moment in the summer of 1962, John Blair, Sen. Kefauver's chief economist, and Lloyd Cutler, a lawyer representing opponents of drug reform, suddenly found themselves agreeing that only substantial—not preponderant—evidence should be required to show the efficacy of new drugs. What moved Blair to abandon the preponderance standard was his belief that when opinions differ, the possibility that truth might reside with the minority must be kept open. In reply to the argument that if a hundred doctors evaluate a drug the opinion of 20 should not prevail, "Blair said that since, generally speaking, any innovation was apt to meet with opposition, he was willing to accept 'substantial' evidence" (Harris 1964, 204). It seems Blair, in the tradition of John Stuart Mill, was wary of establishing a tyranny of the majority that would crush the dissenting views on which progress depends. Cutler, representing a former medical director of the FDA, argued before a House committee that same summer that a drug should be considered effective provided only that responsible clinicians deem it so, even when "a numerical majority or a preponderance of evidence, or whatever else you might call it" points to the opposite conclusion. Somehow the seemingly abstract issue of standards of evidence became tangled up with emotional topics like minority rights and the persecution of dissenting views. Cutler's client spoke powerfully of the evils of orthodoxy and the vital importance of conflicts of opinion, invoking great medical reformers like Semmelweis (Mill's contemporary) in opposition to what he considered the authoritarian strain in proposals for drug reform. But regardless of our cherished stories of lone dissenters braving opposition, it's a strange thought that the preponderance standard fell victim to anti-majoritarian sentiment, just as it would be melodramatic to portray FDA approval of certain New Drug Applications with only 14% positive trials (Fava et al. 2003) as a victory of truth over persecution.

As the story goes, it was Blair who turned compromise into triumph by inserting the all-important reference to clinical trials into the Drug Amendments. In return for scrapping the preponderance standard he "insisted that the amendment

stipulate that the evidence consist 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.'" The deal was done, and when an ally of Kefauver's learned of it, he rejoiced. "I just couldn't believe it when Blair pulled that off," he said. "It gives us all kinds of power—especially the word 'adequate'—to make sure that drugs do what is claimed for them" (Harris 1964, 204–5).

Over the decades since 1962, the FDA has taken the words "adequate and well-controlled investigations" to heart like a motto. An agency endowed with "all kinds of power" by a phrase has become the most redoubtable regulatory body in the world. At this point, however, a curious anomaly appears. While the FDA, empowered by language giving it oversight of clinical investigations of new drugs, has often wielded its regulatory authority with a strong hand, the evidence generated by such investigations is evaluated under a standard designed to be weak, a standard whose very description tends to be watery: "evidence that doesn't necessarily convince everyone." Perhaps giving power with one hand and taking it away with the other was somebody's idea of checks and balances. But if competent investigators produce mixed data as they go about testing whether drugs in fact do what they are represented to do, what becomes of the evidence that tells against efficacy but doesn't stand in the way of a drug's approval, given that the evidence in support of efficacy need only be "substantial"?

When Kirsch (2010) petitioned the FDA under the Freedom of Information Act for its unpublished records of antidepressant trials, he discovered an internal memo contending that the disclosure of such data was "of no practical value either to the patient or the prescriber" (49). It's not hard to see why many might take such a dismissive view of data that looks like a mere by-product or even waste product of the process of drug approval. Overruled by positive findings, the reports of negative trials are filed away—consigned to Lethe. The view that negative findings lack value and therefore do not warrant publication is not, of course, confined to internal FDA discussions. While a recent meta-analysis of antipsychotic clinical trials reaching back to 1960 found that the test drug did better than placebo in only 15 of 39 studies, even this figure may be inflated "because RCTs failing to demonstrate significant differences between medication and placebo may not have been published" (Rutherford et al. 2014, 1418).

With the market overrun with drugs of unverified efficacy at the time of the 1962 Amendments, the need for an appropriate standard of evidence—preponderance—cried aloud. Congress did not heed it. Perhaps fearing that a standard too robust would empty the shelves of the local pharmacy, it "did not intend to set the bar for efficacious new drug approvals too high" (Junod 2008) and therefore imposed a criterion whose only ascertainable meaning is that it is not particularly demanding. If the 1962 Drug Amendments had required a preponderance of evidence from "adequate and well-controlled investigations," the process of

drug evaluation at all levels would have to reckon with contrary evidence more fully and openly than is now the case. Two positive trials would not cancel any number of negative trials, and because negative findings would retain their significance, the argument against their disclosure would lose its foundation. Published literature would not resemble a 1950s medical ad containing strategic omissions. In 1962, however, the suppression of something close to a preponderance of all investigational findings may have been unimaginable.

References

Anello, C., and S. Junod. 2013. "Contributions of Robert T. O'Neill to the Evolution of Regulatory Statistical Science." *Stat Biopharm Res* 5: 1–18.

Avorn, J. 2005. "FDA Standards: Good Enough for Government Work?" *NEJM* 353 (2005): 969–72.

Beecher, H. 1955. "The Powerful Placebo." *JAMA* 159: 1602–6.

Carpenter, D. 2010. *Reputation and Power: Organizational Image and Pharmaceutical Regulation at the FDA*. Princeton: Princeton University Press.

Chalmers, I. 1985. "Proposal to Outlaw the Term 'Negative Trial.'" *BMJ* 290: 1002.

Christakis, N. 2008. "Does This Work for You?" *BMJ* 337: 1025.

Darrow, J. 2013. "Pharmaceutical Efficacy: The Illusory Legal Standard." *Washington Lee Law Rev* 70: 2073–2136.

Fava, M., et al. 2003. "The Problem of the Placebo Response in Clinical Trials for Psychiatric Disorders: Culprits, Possible Remedies, and a Novel Study Design Approach." *Psychother Psychosom* 72: 115–27.

Feinstein, A. 1971. "How Do We Measure 'Safety' and 'Efficacy'?" *Clin Pharmacol Ther* 12: 544–58.

Firth, S. 2015. "FDA Advisory Panel Unmoved by Gepirone Data." *MedPage Today*, Dec. 3.

Goodrich, W. 1963. "A Look into the Nooks and Corners of the Kefauver-Harris Drug Amendments of 1962." *Business Lawyer* 19: 187–92.

Greene, J., and S. Podolsky. 2012. "Reform, Regulation, and Pharmaceuticals: The Kefauver-Harris Amendments at 50." *N Engl J Med* 367: 1481–83.

Harris, R. 1964. *The Real Voice*. New York: Macmillan.

Ioannidis, J. 2017. "Does Evidence-Based Hearsay Determine the Use of Medical Treatments?" *Soc Sci Med* 177: 256–58.

Junod, S. W. 2008. "FDA and Clinical Drug Trials: A Short History." http://www.fda.gov/AboutFDA/WhatWeDo/History/Overviews/ucm304485.htm/.

Kesselheim, A., and J. Darrow. 2015. "FDA Designations for Therapeutics and Their Impact on Drug Development and Regulatory Review Outcomes." *Clin Pharmacol Ther* 97: 29–36.

Khan, A., S. Khan, and W. Brown. 2002. "Are Placebo Controls Necessary to Test Antidepressants and Anxiolytics?" *Int J Neuropsychopharmacol* 5: 193–97.

Kirsch, I. 2010. *The Emperor's New Drugs: Exploding the Antidepressant Myth*. New York: Basic Books.

Kulynych, J. 1999. "Will FDA Relinquish the 'Gold Standard' for New Drug Approval? Redefining 'Substantial Evidence' in the FDA Modernization Act of 1997." *Food Drug Law J* 54: 127–49.

Lasagna, L. 1962. *The Doctors' Dilemmas*. New York: Harper.

Lasagna, L. 1989. "Congress, the FDA, and New Drug Development: Before and After 1962." *Perspect Biol Med* 32: 322–43.

Le Noury, J., et al. 2015. "Restoring Study 329: Efficacy and Harms of Paroxetine and Imipramine in Treatment of Major Depression in Adolescence." *BMJ* 351: h4320.

Martenyi, F., E. Brow, and C. Caldwell. 2007. "Failed Efficacy of Fluoxetine in the Treatment of Posttraumatic Stress Disorder." *J Clin Psychopharmacol* 27: 166–70.

McFadyen, R. 1973. "Estes Kefauver and the Drug Industry." PhD diss., Emory University.

McKinlay, J. 1981. "From 'Promising Report' to 'Standard Procedure': The Seven Stages in the Career of a Medical Innovation." *Milbank Q* 59: 374–411.

Potter, W., C. Mallinckrodt, and M. Detke. 2014. "Controlling Placebo Response in Drug Development: Lessons Learned from Psychopharmacology." *Pharm Med* 28: 53–65.

Rutherford, B., et al. 2014. "Placebo Response in Antipsychotic Clinical Trials: A Meta-Analysis." *JAMA Psychiatry* 71: 1409–21.

Ryan, N. 2005. "Treatment of Depression in Children and Adolescents." *Lancet* 366: 933–40.

Simes, R. 1986. "Publication Bias: The Case for an International Registry of Clinical Trials." *J Clin Oncology* 4: 1529–41.

Spiller, R. 1999. "Problems and Challenges in the Design of Irritable Bowel Syndrome Clinical Trials: Experience from Published Trials." *Am J Med* 107: 91s–97s.

Storosum, J., et al. 2001. "Short-Term Efficacy of Tricyclic Antidepressants Revisited: A Meta-Analytic Study." *Eur Neuropsychopharmacol* 11: 173–80.

Tramer, M., et al. 1997. "Efficacy, Dose-Response, and Safety of Ondansetron in Prevention of Postoperative Nausea and Vomiting: A Quantitative Systematic Review of Randomized Placebo-Controlled Trials." *Anesthesiology* 87: 1277–89.

Turner, E., et al. 2008. "Selective Publication of Antidepressants Trials and Its Influence on Apparent Efficacy." *N Engl J Med* 358: 252–60.

Tuttle, A., et al. 2015. "Increasing Placebo Responses over Time in U.S. Clinical Trials of Neuropathic Pain." *Pain* 156: 2616–26.

Wardell, W., and L. Lasagna. 1975. *Regulation and Drug Development*. Washington, DC: American Enterprise Institute.