Advancing FDA's Regulatory Science through Weight of Evidence Evaluations

Joseph W. Cormier

Follow this and additional works at: http://scholarship.law.edu/jchlp

Recommended Citation
Available at: http://scholarship.law.edu/jchlp/vol28/iss1/2

This Article is brought to you for free and open access by CUA Law Scholarship Repository. It has been accepted for inclusion in Journal of Contemporary Health Law & Policy by an authorized administrator of CUA Law Scholarship Repository. For more information, please contact edinger@law.edu.
ADVANCING FDA’S REGULATORY SCIENCE THROUGH WEIGHT OF EVIDENCE EVALUATIONS

Joseph W. Cormier*

I. INTRODUCTION

The United States Food and Drug Administration ("FDA" or "the Agency") regulates roughly one quarter of the country’s gross domestic expenditure.1 The breadth of products under its regulatory oversight include not only most foods,2 human and veterinary drugs, biologics, medical devices, dietary supplements, and cosmetics, but also radiation emitting products, animal feed, and tobacco products.3 Although FDA has long been recognized as the “gold standard” of science-based regulation of food and

---

1. J.D., Georgetown University Law Center; Ph.D., Columbia University; A.B., Dartmouth College. Dr. Cormier is an associate in the Washington, DC office of Arnold & Porter LLP and is a member of the FDA and Healthcare Practice Group; the views expressed in this article are those of the author and not necessarily those of either Arnold & Porter or its clients. From 2005 to 2011, Dr. Cormier served as a scientist for the U.S. Food and Drug Administration, during which time he focused on the regulation of animal biotechnology. He wishes to thank Lauren Haertlein for her help in the preparation of this article, and the Journal of Contemporary Health Law and Policy staff for all of their efforts.


medical product safety, as the Agency begins its second century, it is striving to be more “science-led.”

Indeed, FDA recently stated in an August 2011 report on regulatory science that:

FDA plays a critical role in protecting and promoting the nation’s health and regulates industries that are among the most successful and innovative in the world. Critical responsibilities across the products FDA regulates require application of the best available science to keep pace with these advances and make decisions and take actions that both support innovation and protect and promote the public health.5

Advancing this effort—fostering an agency whose regulatory approach is driven by science, not just founded on general scientific principles—will require FDA to take a fresh look at regulations and policies that are, in some cases, decades old.6 While some of these regulations and policies may have been sufficient at one time, the authors of some of these older regulations and policies could not have envisioned the types of products or indications that come before the Agency today.7 In other cases, bright-line rules that have been applied over time and have become entrenched fail to adequately address the more nuanced scientific realities of specific applications. In each case, FDA has in good faith established a regulatory paradigm founded on sound science. However, when application-specific scientific data suggest that FDA regulate in a more individualized manner, a science-led agency should be free to do so.

One example of this issue is in how FDA determines that a new drug or biologic is effective in meeting its claims and intended use. The Federal


5. U.S. FOOD & DRUG ADMIN., supra note 1, at 35.

6. See, e.g., 21 C.F.R. pt. 211 (2011) (good manufacturing practices for drug products). Although Part 211 was first finalized in 1978, 43 Fed. Reg. 45,077, only minor changes have been made since the mid-1980s. Id.

7. For example, the use of modern biotechnology to genetically engineer animals to produce human biologics.
Food, Drug, and Cosmetic Act ("FDCA") requires that sponsors provide "substantial evidence" demonstrating that their products are effective under the prescribed conditions for its intended use. This article will first review the statutory requirement and FDA's historical interpretation of "substantial evidence." Second, the article will argue that a rigid application of FDA's interpreted standard can lead to scientifically inappropriate results. Third, the article will: (1) discuss an alternative approach that utilizes a weight of evidence evaluation when determining the sufficiency of scientific data and information, and (2) examine two general examples of how such an approach would be beneficial to FDA, drug and biologic sponsors, and the general public. Finally, the article will consider some of the policy considerations that must be taken into account prior to implementing a weight of evidence approach to data review.

II. THE "SUBSTANTIAL EVIDENCE" REQUIREMENT FOR CLINICAL EFFECTIVENESS

Manufacturers of drugs and biologics are required to obtain a positive affirmation of both safety and effectiveness of their products from FDA in the form of a product approval prior to initial marketing. In reviewing such applications, FDA ultimately decides if a product's benefits outweigh its risks.

A. Drugs

It has been observed that "[t]he history of clinical trials closely follows the history of drug regulation." The first federal regulation of drug products, the Pure Food and Drug Act of 1906, prohibited the sale of misbranded or


9. FDCA § 505(a), 21 U.S.C. § 355(a) (prohibiting the introduction into interstate commerce of any drug that has not been approved); Public Health Service Act § 351, 42 U.S.C. § 262(a) (2006) (prohibiting the introduction into interstate commerce of any biologic that has not been approved).


adulterated pharmaceuticals.\(^\text{12}\) In 1938, after more than one hundred people died from using an unsafe drug product, Congress passed the FDCA.\(^\text{13}\) With the 1938 FDCA, FDA only assessed product safety. This changed in 1962 when concerns about misleading and unsubstantiated product-claims led Congress to amend the FDCA to include an efficacy requirement.\(^\text{14}\) Drug efficacy is evidence of a causal relationship between the product and a clinical benefit. Researchers and pharmacologists had testified before the Senate about the importance of quality clinical trials in pharmaceutical market control.\(^\text{15}\) The 1962 Kefauver-Harris Amendments to the FDCA incorporated these concerns into the statutory requirements for FDA market approval.\(^\text{16}\)

Under the amended FDCA, sponsors must provide "substantial evidence" demonstrating that their products are effective under the prescribed conditions for their intended use.\(^\text{17}\) The FDCA defines "substantial evidence" as:

* 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 responsibly be concluded by such experts that the drug

---


16. See id. at 134 n.48 (discussing expert testimony on effectiveness requirements); *FDA History-Part III: Drugs and Foods Under the 1938 Act and Its Amendments*, U.S. FOOD & DRUG ADMIN., http://www.fda.gov/AboutFDA/WhatWeDo/History/Origin/ucm055118.htm (last updated June 18, 2009).

will have the effect it purports or is represented to have under the conditions of the use prescribed, recommended, or suggested in the labeling or proposed labeling thereof.\textsuperscript{18}

Initially, FDA interpreted “investigations” to mean at least two human studies were required.\textsuperscript{19} Over time, however, FDA permitted the use of a single study on a case-by-case basis. The Food and Drug Administration and Modernization Act of 1997 (“FDAMA”) codified this practice, amending the FDCA to explicitly allow for the use of one clinical study to support substantial evidence under certain limited circumstances.\textsuperscript{20} FDAMA specifically provides that FDA may consider “data from one adequate well-controlled clinical investigation and confirmatory evidence” to constitute “substantial evidence” if FDA determines, “based on relevant science,” that the data establishes effectiveness.\textsuperscript{21}

\section*{B. Biologics}

The Public Health Service Act (“PHSA”) governs the FDA regulatory approval for biologic products.\textsuperscript{22} Under PHSA, FDA may approve biologics once they have been demonstrated to be “safe, pure, and potent.”\textsuperscript{23} In FDAMA, Congress instructed FDA to minimize the differences in the review and approval of biologics under the PHSA and drugs under FDCA.\textsuperscript{24} Accordingly, FDA has historically interpreted “potency” to mean effectiveness and incorporates the new drug application standard for

\begin{itemize}
\item \textsuperscript{18} Id. (emphasis added).
\item \textsuperscript{19} \textit{E.g.}, Warner-Lambert Co. v. Heckler, 787 F.2d 147, 151 (3rd Cir. 1986). The term “clinical” in the definition is routinely interpreted to mean “human.” \textit{See, e.g.}, Russell Katz, FDA: Evidentiary Standards for Drug Development and Approval, \textit{1 NEURORx} 307, 308 (2004).
\item \textsuperscript{21} Id. § 115(a).
\item \textsuperscript{23} Id.
\item \textsuperscript{24} FDAMA § 123(f).
\end{itemize}
adequate and well-controlled studies to demonstrate effectiveness. Therefore, the standards for substantial evidence are essentially identical for drugs and biologics.

C. Historical Interpretation

Since the 1962 amendments to the FDCA, FDA, the pharmaceutical industry, and the scientific community at large have debated what constitutes sufficient evidence of effectiveness. Congress tasked FDA with specifying the legal and scientific evidentiary requirements by leaving the phrase “adequate and well controlled” undefined. FDA has therefore promulgated regulations that detail “characteristics . . . recognized by the scientific community as the essentials of an adequate and well-controlled clinical investigation.” These characteristics specify the criteria for an “adequate and well-controlled” study: (1) a pre-determined study objective and method for the analysis of results; (2) a design that allows one to validly and quantitatively compare the test drug to a control (usually either another drug or placebo); (3) a method for selecting appropriate subjects; (4) a method for assigning treatments; (5) methods for minimizing bias from subjects, observers, and data analysts; (6) methods for assessing subjects’


26. Throughout the remainder of the article, unless otherwise noted, the terms “drug” and “biologic” will be used interchangeably.


30. 21 C.F.R. § 314.126(a).
responses; and (7) a description and analysis of results and methods of evaluation, "including any appropriate statistical methods." 31

When finalizing this interpretation, FDA stated that the regulation has two objectives: (1) to minimize bias in clinical studies, and (2) to assure that the study methods are sufficiently detailed so that FDA can fully assess and interpret the study data. 32 Therefore, the goal of adequate and well-controlled studies is to produce valid and reliable data and reduce the likelihood that an observed benefit is something other than a drug effect. 33 When assessing the core questions of safety and effectiveness, FDA takes an empirical approach. 34 Regulations "require a quantitative comparison of the effects of the drug to . . . a control group," 35 but do not require any particular method of statistical analysis when considering the results. 36

III. CLINICAL EVALUATIONS AND STATISTICAL ANALYSIS

FDA's general practice is that applications must include two trials with significant results. 37 However, after FDAMA, FDA published a guidance document detailing the regulatory and scientific considerations for approving a marketing application based on a single "adequate and well-controlled" study. A single study may be sufficient when related data can substantiate the study—such as new dose regimes for approved products—and when a single multicenter trial provides evidence of effectiveness, supported by confirmatory research. 38

31. Id. § 314.126(b)(1)-(7).


33. See, e.g., Katz, supra note 19, at 309-10 ("The Agency routinely seeks to minimize the likelihood that any beneficial effect seen in a drug trial is the result of [fraud, bias, or chance].").

34. Id. at 316.

35. Id. at 311.

36. Id.


38. GUIDANCE ON EVIDENCE, supra note 27, at 3-4.
Generally, reliance on a single study is based on an "extreme P-value." FDA has delineated other exceptions to its statistical paradigm, including regulations on the use of surrogate endpoints. Restricting the application of these modified evidentiary requirements to serious illnesses, however, "reflect[s] the Agency's acknowledgement that these approvals introduce a level of uncertainty into the approval process that is ordinarily not present (namely, the uncertainty that the effect of the drug on the surrogate will predict the desired clinical benefit)." As opposed to standard outcome measures like increased survival, surrogate endpoints are only presumed to predict desired clinical effects. For example, tumor shrinkage might be reasonably likely to predict that a product effectively extends cancer patients' survival. FDA's regulations permit the Agency to approve products for serious or life-threatening diseases based on clinical trials that establish that "the drug product has an effect on a surrogate endpoint that is reasonably likely, based on epidemiologic, therapeutic, pathophysiologic, or other evidence, to predict clinical benefit."42

FDA has stated that "[a]lthough there is no statutory requirement for significance testing of any particular value, there are well-established conventions for assessing statistical significance to support the statutorily required conclusion that the well-controlled studies have demonstrated that a drug will have the effect it is represented to have."43

39. E.g., Robert Temple, How FDA Currently Makes Decisions on Clinical Studies, 2 CLINICAL TRIALS 276, 277 (2005). The statistical probability that an observed difference between treatment groups results from chance is represented by the p value.

40. See 21 C.F.R. § 314 Subpart H—Accelerated Approval of New Drugs for Serious or Life-Threatening Illnesses (2011); see also 21 C.F.R. § 601 Subpart E—Accelerated Approval of Biological Products (2011).

41. Katz, supra note 19, at 309.

42. 21 C.F.R. § 314.510.

Typically, FDA requires a statistically significant difference in some measure of clinical signs or symptoms between the treatment and control groups. \(^{44}\) Significance in this context is how likely it is that a difference as large as the one observed occurred purely by chance. Results are generally considered “significant” when the probability of making a false-positive claim about the beneficial effect of a treatment is less than five percent. \(^{45}\) A \(p\) value of less than 0.05 indicates a less than five percent probability that an observed drug-benefit occurred by chance. \(^{46}\)

FDA generally considers a clinical trial to be a success if the \(p\) value is less than or equal to 0.05 when comparing the treatment group to the control group. \(^{47}\) For example, Dr. Robert Temple, Director of FDA’s Office of Drug Evaluation I, remarked to an FDA advisory panel that “substantial evidence” of a “clinically meaningful effect” generally “means at least one study showing a very large effect, or two studies for which the \(P\) value is less than 0.05.” \(^{48}\) Although FDA’s interpretation of the characteristics of an adequate and well-controlled study were not intended to be viewed as a

\(^{44}\) See, e.g., Lawrence Gould, Substantial Evidence of Effect, 12 J. Biopharmaceutical Stat. 53, 54 (2007) (describing the traditional application of “substantial evidence” as two trials demonstrating that the product has a statistically significant effect).

\(^{45}\) See Henry I. Miller & David R. Henderson, The FDA’s Risky Risk-Aversion, 145 Pol’y Rev. 14, 15-17 (2007) (“Although arbitrary, the bar is typically set at 5 percent . . . as ‘proof for most phenomena in the realms of medicine and science.’”).


\(^{47}\) E.g., Katz, supra note 19, at 310; see also Miller & Henderson, supra note 45, at 15–17 (discussing multiple examples of when FDA delayed or denied approval for drugs that despite other evidence, were not shown to be statistically significant at primary endpoints, and discussing an example of when FDA required additional studies to demonstrate effectiveness when initial studies fell short of the \(p \leq 0.05\) threshold).

checklist, FDA “almost universally” applies this standard to its efficacy evaluations. Accordingly, what generally constitutes “substantial evidence” in practice are results that demonstrate drug-effect at a five percent level of significance.

The measure of statistical significance being at the five percent level is more or less an artifact of historical chance, when a full range of statistical tables were difficult to publish due to the sheer number of tables required. As a result, R. A. Fisher’s seminal 1925 text on the subject, although providing tables with multiple levels of significance for other values, only provided the five percent level for one particular table. This value subsequently became the standard of significance for the biological and medical sciences.

Statistical approaches to data analysis presuppose the existence of confounding variability within and between sets of data. Statistics is attempting to solve the difficulty of searching for “scientific truth” in “randomness.” Ultimately however, statistical significance is merely an arbitrary line along a spectrum of probabilities. When statisticians say that the statistical significance of a given observation has a $p$ value of 0.05, what they are saying is that there is a one-in-twenty chance that the observed difference is a product of random variability within a data set. Said differently, if there truly was no difference between test articles and the same experiment were conducted twenty times, we would expect to falsely “find” a difference once. Statistical analysis, then, is merely a quantified statement of how confident one is in the conclusions drawn from data. Statistical significance is placing a bright line along that spectrum in order to produce a binomial decision—yes there is a difference, or no, there is not.


50. Katz, supra note 19, at 311.


52. See Stephen Stigler, Fisher and the 5% Level, 21 CHANCE 12, 12 (2008) (discussing the development of 5% as an adequate measure of significance).

53. See id.

54. Id.
Drawing these lines of distinction renders the conclusions of significance both under-inclusive and over-inclusive. First, observations that clearly have practical meaning may fall short of statistical significance due to the statistical power of a given study. For example, if one drug in a given class demonstrates effectiveness with a \( p \) value of 0.04 after a very large clinical trial, and a second drug within the same class—and as to which all scientific principles suggest would act similarly—demonstrates effectiveness with a \( p \) value of 0.06 after a smaller study, it would violate reason to say that the first is effective whereas the second is not. It would also be an inefficient use of resources (and potentially unethical) to force the sponsor of the second drug to recruit additional subjects when the result of lowering the \( p \) value to reach 0.05 is more or less a foregone conclusion. Second, data analysis may show that a statistical significance exists when such a significance has no meaning in practice. For example, a clinical study for a topical antibiotic ointment may show that individuals given the treatment, as opposed to a placebo, had a small, but statistically significant increase in the development of gastric ulcers. Given that there is no reason to expect that local, topical application of an antibiotic would have any causal relationship to ulcers, it should be unnecessary to conduct a full follow-up study to demonstrate the lack of such a relationship, particularly when other similar medications are already known not to have such an effect. In each of these cases, the statistical analysis fails in that it becomes divorced from basic first principles of science.

IV. THE WEIGHT OF EVIDENCE ALTERNATIVE

One potential answer to the problems associated with rigid application of a bright-line statistical rule is a weight of evidence analysis. A weight of evidence approach to data analysis allows the decision-maker to look at all data and information, whatever its value, and give each its proper consideration. For example, this would allow a reviewer to consider a study whose data demonstrate a statistical \( p \) value that, while not technically meeting a standard definition of "significance," nonetheless provides evidence of safety or effectiveness. It is important to note that such an approach does not abandon statistical analysis, but rather borrows from the Bradford Hill criteria for causation when considering the question of whether the data are indicative of real differences.

Bradford Hill's criteria for causation require an individual to consider nine specific aspects of a given observation when attempting to determine the

---

linkage between the data and the proposed cause. First, although many very real and consequential effects are the result of small differences between a treatment group and a control group, one should give consideration to the strength of the association between the data and the hypothesis: the larger the effect, the stronger the association. Second, one must consider the consistency with which a given observation is made. This consideration is a measure of how robust a given conclusion is with respect to the experimental question. Third, the specificity of the association is evaluated. This evaluation is a measure of how unique a given association is relative to other observations. The fourth consideration is the temporal nature of the observation and the potential cause. Fifth, one must evaluate the dose-response relationship, that is, the relationship between the strength of the association and the degree of exposure to the alleged causal agent. Sixth is that of plausibility, whether the association is congruous with the scientific possibility of a causal link. Seventh, and related to plausibility, is coherence; causal associations should not be incoherent with generally known scientific facts. Eighth, one may wish to look at an experiment intended to evaluate the alleged causal relationship. This step, however, brings with it all of the various difficulties already mentioned and must, in turn, be evaluated based on the other Bradford Hill criteria. Finally, one can consider whether analogical reasoning supports the proposed causal association.

56. See generally id.
57. Id. at 295-96.
58. See id. at 296.
59. See id. at 297.
60. See id.
61. See Bradford Hill, supra note 55, at 298.
62. See id.
63. See id.
64. See id.
65. See id. at 299.
Taken together, when reviewing an individual set of data and the question of causation, the reviewer should look at the strength of the association (the statistical analysis) in the context of the data's internal consistency as well as its coherence with first principles of science and biological plausibility.

Such an approach to data analysis is not new to FDA or to other government agencies. The Environmental Protection Agency regularly utilizes a weight of evidence approach to determine acceptable levels of various substances in drinking water and the atmosphere. The FDA also regularly invokes the weight of evidence concept when communicating issues of causation, for example, when considering the toxicity of a regulated product or a qualified health claim for a food. In a 2009 briefing on the status of FDA regulatory science, the Agency stated that "regulatory and public health decisions promulgated by the FDA are based upon the weight of scientific evidence.” Nonetheless, FDA rarely articulates what it means when it says “weight of evidence.” In one instance, FDA’s Center for Veterinary Medicine (“CVM”) has provided some detail regarding the Agency’s approach to weight of evidence analysis by stating that FDA “draw[s] on data from a number of sources” including controlled studies on target populations, non-controlled studies on target populations, and other studies in the available scientific literature that either involve the specific product and target population or are related to the product under consideration. Importantly, “[i]rrespective of the source or [level] of

---


deference given to a given dataset, all of the data and information is evaluated in the context of basic scientific principles and external validity. 69

Notably, CVM used a weight of evidence approach to approve a new animal drug application for ATryn®-producing goats in 2009 without the sponsor submitting a single effectiveness study that would have traditionally been required under the “substantial evidence” standard. 70 In this case, several factors led FDA to the conclusion that the data nonetheless amounted to substantial evidence of safety and effectiveness. First, the nature of the genetic construct used (the drug article) suggested that the article would be expected to be present only in a localized part of the animal. Second, the sponsor provided historical records demonstrating overall herd health that was consistent with traditional herds. 71 Finally, because the effect (the presence of the human biologic in the milk of the goats) is not logically expected to be present in non-engineered goats, no blinded study was required; the substantial evidence of effectiveness was established by simply demonstrating that human antithrombin III was present in the goats’ milk. 72 Based on the weight of evidence, as characterized by the data provided in the context of scientific plausibility, CVM concluded that the data met the substantial evidence standard statutorily required by the FDCA. 73 Therefore, FDA was on solid footing when issuing its approval for the New Animal Drug Application (“NADA”), and to date, has not been challenged on that finding. The statutory standards for safety and effectiveness for NADAs are essentially identical to human new drug applications (“NDAs”) 74 (and in practice, the same for biologic licensing applications (“BLAs”) as well).

69. Id. (emphasis added).


71. This did not include a traditional study, but rather used the broader scientific and clinical knowledge of general goat herd health, as determined by examination of published literature as well as the experience of trained veterinarians and animal scientists. Id.

72. Id.

73. Id.

Therefore, it is possible that the same set of circumstances—where scientific principles lead to the conclusion that the traditional blinded studies do not add to the regulatory decision-making process—could arise within the context of a human drug or biologic application, and, if so, a similar conclusion would be just as scientifically and legally sound.

A. Weight of Evidence Evaluations in Practice

While a weight of evidence evaluation generally appeals to logic and the scientific process, how it is implemented in practice is less clear. A scientific reviewer at FDA is still tasked with deciding whether there is substantial evidence that a given product is effective for its proposed indication. Turning the “analog” weight of evidence evaluation into a “digital” regulatory conclusion requires significant expert scientific judgment. This subjectivity can lead to a lack of clarity or transparency as to how a regulatory decision-maker arrived at his or her conclusion. It is the rationale of these decisions that informs future applications and “fills in” the contours of otherwise ambiguous statutory language.

Therefore, without a general framework to guide the regulatory scientist, FDA risks the appearance of making arbitrary decisions and sponsors are left without a full understanding of the expectations for their submissions. A weight of evidence evaluation, however, need not be a rudderless ship. As is outlined in the following Table, FDA has provided some broad strokes of how it conducts such reviews.

75. A weight of evidence review allows FDA staff to determine where along a broad spectrum of evidence the sum total of support for an application lies. This process, therefore, can conclude with any one of an infinite number of results between “no data or information” and “scientifically irrefutable fact.” Irrespective of how data and information are reviewed by FDA staff, the Agency is ultimately charged with determining whether or not a given application will be approved. To put it perhaps too simply, FDA has one of two choices when considering whether to allow a product to be marketed: “yes” or “no.”
GENERAL CONSIDERATIONS FOR WEIGHT OF EVIDENCE EVALUATIONS

BASIC PRINCIPLES OF SCIENCE
(BIOLOGICAL PLausIBILITY AND COHERENCE)

EXTERNAL VALIDITY

<table>
<thead>
<tr>
<th>Order of</th>
<th>Description</th>
<th>Considerations</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deference</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>• Controlled Studies</td>
<td>• Study Quality</td>
<td>• Large, Double Blind</td>
</tr>
<tr>
<td></td>
<td>• Same Animal Species</td>
<td>• Relevance of Endpoint</td>
<td>• Use of “Good Study Practices” such as Good Clinical Practices</td>
</tr>
<tr>
<td></td>
<td>• Internal Validity</td>
<td>to Regulatory Question</td>
<td>• Full Data Set</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Agreed-to Study Design</td>
</tr>
<tr>
<td>2</td>
<td>• Non-controlled Studies</td>
<td>• Study Size / Duration</td>
<td>• Pilot Study</td>
</tr>
<tr>
<td></td>
<td>• Same Animal Species</td>
<td>• Study Quality</td>
<td>• Very Small</td>
</tr>
<tr>
<td></td>
<td>• Same Regulated Product</td>
<td></td>
<td>• Different Endpoint</td>
</tr>
<tr>
<td>3</td>
<td>• Historical Summary</td>
<td>• Study and Data Quality</td>
<td>• Summary Data Only</td>
</tr>
<tr>
<td></td>
<td>• Same Animal Species</td>
<td></td>
<td>• No Study Design or Design Not Followed</td>
</tr>
<tr>
<td></td>
<td>• Same Regulated Product</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• “Epidemiology” Study</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>• Studies on Similar or Related Animal Species</td>
<td>• Degree of Similarity to the Regulated Product and the Target Species</td>
<td>• Different Animal Species</td>
</tr>
<tr>
<td></td>
<td>• Different Regulated Product, but Potentially in Same or Similar Class</td>
<td></td>
<td>• Different Regulated Product that is Expected to Have the Same or Similar Effect</td>
</tr>
</tbody>
</table>

76. Adapted from U.S. FOOD & DRUG ADMIN. CTR. FOR VETERINARY MED., supra note 68, at 3.
When using this approach, an FDA review scientist begins his or her evaluation of a piece of data or information by considering the various qualities of the study presented. This “study” may range from the traditional double-blinded, placebo-controlled study to historical summaries of related information (even potentially using a different regulated article) in an entirely different species of animal (for example, a rabbit study using a drug in the same class as the drug under consideration in the application). The reviewer considers the various Bradford Hill criteria discussed above, paying particular attention to the strength and specificity of the association, as well as the robustness of the study, which would include the evaluation of study bias. Depending on this evaluation, a reviewer will then be able to place it along a continuum of deference, from which just four different points are illustrated in the Table. Importantly, each of the considered data points is evaluated in the overarching context of biological plausibility and coherence as well as whether the data are strengthened by external validity, which is to say whether the data are independently substantiated by other data or information.

B. Case Studies

1. Review of Long-Adopted Medical Claims

Weight of evidence evaluations could be employed where the medical community has already accepted the demonstrations of safety and effectiveness of a specific product claim. In such circumstances, it may be impractical to conduct clinical trials because of the difficulty in recruiting physicians to give at least some of the enrolled study subjects the placebo control.

One example is the prophylactic treatment of the eye with a topical antibiotic for the purpose of reducing the potential for surgical-related infection. Well over one million cataract surgeries are performed in the United States each year. Endophthalmitis is a relatively rare infection of the eye that can, in severe cases, result in significant vision loss even if it is appropriately treated. Numerous clinical studies reported in the scientific


literature as early as 1952 have shown a marked decrease in the incidence of post-operative endophthalmitis associated with the prophylactic treatment of antibiotics after ophthalmic surgery. Although there is some recent debate over whether such treatments are more or less responsible for observed reductions in incidence rates (as opposed to other factors such as sterile surgical methods), even critics nonetheless recommend prophylactic antibiotic use in combination with other infection-reducing methods because the overall risk posed by such use is far outweighed by the risks associated with endophthalmitis. Clinicians have long stated that “[t]he rationale for such prophylaxis . . . [is] well founded.”

Nevertheless, no ophthalmic topical antibiotic is currently approved in the United States for peri- and post-operative use to prevent endophthalmitis. Due to the potential risks associated with the disease, it is not surprising that ophthalmologists are not willing to subject their patients to the chance that they would be given the placebo rather than the antibiotic treatment. This unwillingness is particularly evident when one considers the number of patients who would need to enroll in order to demonstrate an effect when the overall incidence of the disease is so low to begin with. Given that there are no known risks associated with the use of prophylactic antibiotics (other than the potential for increased bacterial resistance), and that the vast majority of U.S. ophthalmologists have successfully used the treatments for decades (and are likely to continue to use them even in the absence of FDA approval, thereby negating any decrease in bacterial resistance as a result of not approving the indication), a weight of evidence approach would allow FDA to consider all of the known risks and benefits without subjecting patients to a potential increase in endophthalmitis during a prospective

79. See, e.g., Michael B. Starr, Prophylactic Antibiotics for Ophthalmic Surgery, 27 SURV. OPHTHALMOLOGY 353, 354 (1983) (Table 1, part B “Series Using One or More Prophylactic Antibiotics by One or More Routes” versus part A “Series Not Using Prophylactic Antibiotics”).

80. See Christopher N. Ta, Minimizing the Risk of Endophthalmitis Following Intravitreous Injection, 24 RETINA 699, 702 (2004) (discussing various studies that support the use of prophylactic antibiotics and others that suggest that it is no more clinically effective than iodine and sterile surgical procedures).

81. Id. at 702-03 (“Despite the controversy surrounding endophthalmitis prophylaxis, . . . [t]he use of topical antibiotics has been shown to reduce conjunctival and eyelid bacterial flora, which main in turn also decrease the risk of endophthalmitis.”).

82. See, e.g., Starr, supra note 79, at 353.
clinical study. Should such an analysis have concluded that additional studies are needed, then sponsors, physicians, and patients would be no worse off than they are presently. However, if the analysis were to find that substantial evidence of safety and effectiveness exists when considering the benefits and risks associated with the intended conditions of use, then the public is served by having treatments approved for a given therapeutic use rather than resorting to off-label use by physicians.

2. Review of New Technology

In addition to areas where a body of evidence is readily available to support an application, a weight of evidence evaluation is also useful in considering new technologies for which a “one size fits all” approach is not appealing. In fact, as discussed with regard to CVM’s review of genetically engineered animals, these circumstances are when FDA is more likely already to have utilized such an approach. As FDA begins to consider how it will approach the review of biosimilars, it has suggested a review paradigm that sounds very similar to the weight of evidence approach detailed here.

The issue of regulating biosimilars is not as straightforward as one might think. Unlike their small molecule counterparts, biologics are much more difficult to characterize fully and manufacture consistently. Furthermore, slight changes in a biologic’s chemical structure, which may or may not be detectable and which may be impacted by slight variations in the manufacturing process, can have a dramatic impact on a product’s effectiveness.


84. Id.

85. See Steven Kozlowski et al., Developing the Nation’s Biosimilars Program, 365 NEW ENG. J. MED. 385, 385-87 (2011) (noting that “complex structures of biologic products are usually not easily characterized” and that “the manufacturing processes may introduce potential variants or impurities that could affect risk”).

86. See id. at 386 (stating that biologics “must have a specific set of structural features . . . essential to their intended effect, and slight modifications . . . [or] inadvertent chemical modifications” can affect their performance or safety profile).
When discussing its intended approach to dealing with these issues, FDA states that the administration "scientists will need to integrate various types of information" when making their regulatory decisions.\(^8\) In considering FDA's historical data analyses, the Agency notes that integrating different pieces of data and information when considering product applications is a part of its traditional process.\(^8\) FDA has proposed using a "totality of the evidence" approach that evaluates product attributes "with multiple complementary methods."\(^9\) Recognizing that requiring sponsors to conduct studies that are unnecessary is unethical, FDA proposes utilizing data and information that already exists\(^9\) to ensure that any required studies are in addition to that knowledge and are "carefully tailored to address residual uncertainty" alone.\(^9\)

In sum, conducting a weight of evidence evaluation requires scientific expertise and judgment, but it enables regulatory decision-makers to consider and give weight to a broader range of data, including information that might otherwise fail the traditional, yet somewhat arbitrary, definitions of statistical significance.

V. OTHER ISSUES TO CONSIDER

Although a weight of evidence approach frees FDA review staff from the mandatory application of statistical significance evaluation, some legal and policy issues must be considered before advocating for wide application of this approach.

First, from a legal perspective, the weight of evidence approach has the potential to implicate intellectual property concerns. On its face, because no data is pre-determined to be excluded under a weight of evidence evaluation, a reviewer should consider all data relevant to the question at hand. This consideration potentially includes data in the application and available from other sources—including both public scientific literature as well as data from other applications that would otherwise be considered proprietary. As discussed below, this concern could be addressed by explicitly prohibiting

---

87. See id.

88. See id.

89. Id.

90. See id. at 385.

91. See Kozlowski et al., supra note 85, at 386.
the inclusion of confidential data absent appropriate authorization. However, this alternative is not that different than the current regulatory paradigm in that the administrative record of decisions cannot be based in whole or in part on information that is unavailable to the reviewer by virtue of being proprietary.

Second, although a weight of evidence standard is a different—though not necessarily a less rigorous—approach to regulatory decision-making when compared to the evaluation of statistical significance, few individuals fully understand the limitations of statistical analyses. As a result, most of the public view statistical significance as representing the "gold standard" of causal analysis, and therefore any deviation from that approach represents a lowering of the standards of evaluation. Unfortunately, the complexity of statistical and weight of evidence analyses renders them difficult to fully explain in easy-to-digest terms. Correcting errors in public perception about a change to a weight of evidence standard is likely to be met with broad statements that rouse fear of risks associated with a true lowering of the regulatory standard.

Finally, from a policy perspective, a bright-line rule, such as the traditional approach to the substantial evidence standard, provides little discretion on the part of the decision maker. Once a reviewer agrees that required clinical trials have been properly designed and executed, the statistical analysis of the resulting data is the primary driver of the end decision. Although, as noted, this approach carries distinct disadvantages, a weight of evidence approach would incorporate some additional subjectivity, and potential unpredictability, to the application review process. Therefore, to the extent that an individual is politically disinclined to give FDA reviewers additional discretion, there is likely to be resistance to a weight of evidence approach.

VI. CONCLUSION

As noted above, FDA already uses weight of evidence approaches in many contexts, including in aspects of the evaluation of drugs under the existing statutory standards for safety and effectiveness. However, given the current statutory limitations on the interpretation of the "substantial evidence" standard, and the long history of FDA regulatory interpretations of the standard, a shift to the weight of evidence standard as the general approach to drug approval would likely require a reframing of the definition of "substantial evidence" or a rewriting of the standard altogether.

Adopting language by statute or regulation that explicitly allows for weight of evidence consideration of data has the potential to free FDA from erroneous criticisms that it would be lowering its standards for safety and effectiveness. Such language would also be helpful in avoiding any resistance within the Agency that might incorrectly interpret existing
statutory provisions and regulations defining substantial evidence as requiring rote statistical analyses. Finally, such language could be used to reinforce existing confidentiality rules to ensure that intellectual property, in the form of data submitted under existing applications, is not used in a weight of evidence evaluation of other products unless relevant exclusivity and patent protections have expired.

In conclusion, a weight of evidence approach is one alternative to the current review paradigm at FDA that allows additional data and information to be considered during product review. The Agency itself has noted that:

[FDA is] at a critical moment where advances in science are leading toward fundamental changes in the way medical treatments and diagnostics are being developed and used . . . . Regulatory science must be one step ahead to equip FDA with the necessary tools and methods to reliably assess the safety and efficacy of products derived from these new scientific developments, in order to bring the rewards of discovery safely forward to benefit patients.92

Together, considering any and all available data relevant to a given product as well as refraining from rote application of statistical analysis, FDA can further its stated goal of minimizing bias and allowing a scientific assessment and interpretation of scientific data when evaluating new product applications. By giving review staff the discretion to review applications using a weight of evidence approach, they are given—in fact, encouraged to exercise—the flexibility to consider the application in the broader scientific context. This flexibility, in turn, will lead to a more scientifically-driven analysis that will ensure the advancement of FDA regulatory science and the overall public health.