Evaluating Drug Effects in the Post-Vioxx World
There Must Be a Better Way

Jerry Avorn, MD

The drug approval process must determine efficacy validly, detect risks prudently, and do both in a timely and efficient way. Several high-profile medication withdrawals in recent years have refocused attention on the difficulty of the US healthcare system in meeting these goals. Rofecoxib (Vioxx, Merck) was taken off the market in September 2004 after 5 years of use by ≈20 million people, when its manufacturer reported that the drug doubles the risk of myocardial infarction or stroke. That experience, as well as other instances in which important side effects came to public attention long after a drug’s approval, has prompted discussion about how to repair the nation’s drug evaluation process. Some have proposed changing the degree of statistical certainty required for initial drug approval, with the hope that demanding a probability value of <0.0001 for efficacy (rather than the conventionally required 0.05) would slow down the approval process enough to increase the likelihood that adverse events would be detected. This idea is evaluated and rejected by Roth-Cline in the this issue of Circulation.1 It is true that demanding unprecedentedly stringent tests for efficacy is a poor way to fix the nation’s drug safety problem, but not necessarily for the reasons Roth-Cline advances. A better solution will have to be based on a more comprehensive understanding of how the risks of new drugs are detected, evaluated, and addressed.

Wrong Lesson 1: Preapproval Studies Provide the Only Opportunity to Detect Important Drug Risks

The “extreme certainty of efficacy” proposal cited by Roth-Cline is based on the assumption that preapproval trials are the only setting in which we can feel confident that drug risks will be tracked carefully, a concept that is both true and false. It is false because a well-functioning system of postmarketing safety surveillance could make it possible to detect important adverse effects of a drug, even rare ones, soon after it comes into routine use. If analyzed correctly, use of a new product by hundreds of thousands or millions of typical patients can readily address the power considerations that would otherwise bedevil a modest-sized preapproval trial.3 A robust, rigorous, and efficient system of postmarketing surveillance could be put in place for a tiny fraction of the nation’s $200 billion annual drug expenditure.

On the other hand, the draconian probability value proposal has been advanced in part because the United States does not have such a system of safety surveillance.24 The absence of a coherent national approach to this problem has led some to think that slowing down preapproval efficacy studies is the best tool we have to define drug risks in a timely way. Other trends in risk detection at the FDA also help to explain the appeal of this extreme plan.

In the early 1990s, an implicit social contract was proposed in which the FDA would approve new products more quickly and concurrently develop a better system of monitoring for adverse events once the drugs were in routine use to detect problems that the new speedier approval process may have overlooked. The first part of the plan was implemented with the help of the 1992 Prescription Drug User Fee Act, allowing pharmaceutical companies to pay the FDA to cover the cost of the additional agency staff required to review new drug applications rapidly.4 The time required for new-product approval dropped sharply, but the original Prescription Drug User Fee Act legislation prohibited any application of user-fee funds to support safety studies of those drugs once they were on the market. This exacerbated existing funding problems at the agency’s Office of Drug Safety and culminated with the famous statement by FDA scientist David Graham at a 2004 Senate hearing that the FDA is not capable of protecting the American people from unsafe drugs.
Wrong Lesson 2: Either We Can Approve Drugs in a Timely Manner or We Can Learn Enough About Their Risks Before Approval, but We Cannot Do Both

This flawed assumption might be called the “Heisenberg fallacy.” Such a measurement paradox may apply to electrons but not to drugs. Defenders of this view argue that improving the capacity of preapproval studies to detect risks will inevitably delay the availability of new products and drive up their costs. In using this argument to reject the extreme efficacy evidence proposal, Roth-Cline arrives at the right conclusion for the wrong reason. Drug evaluation is not a zero-sum game requiring impossible tradeoffs among numbers of subjects, probability values, and time, as Roth-Cline suggests. What matters far more is that studies be designed primarily to generate scientific knowledge, not just to grease the wheels toward rapid approval or expansion of a product’s uses. At present, the design of many preapproval trials is more of a problem than the probability values used to evaluate them.

In response to decades of pressure to make the FDA more industry-friendly, we have drifted toward a lowest-possible-standard approach. Often, a manufacturer must show merely that its new product works better than a comparator (often placebo) in achieving a surrogate outcome (eg, improvement in a laboratory test) in a modestly sized sample of atypically younger and healthier patients (compared with expected users of the marketed drug) who are observed over a brief period of time (weeks or a few months, even for medications designed to be used chronically). There are several ways these standards could be made more relevant to practice without dramatically increasing the duration or cost of drug evaluation and without recalibrating the level of statistical certainty required for approval.7

For toxicities that do not depend on the duration of drug exposure, risk detection can be enhanced without extending trial length simply by enrolling more adequate numbers of patients. An even better way to identify potential safety problems would be to correct the present maldistribution of age and comorbidity among study subjects. More than a decade ago, our group reported on the systematic exclusion of older patients in pivotal drug trials in cardiovascular disease.8 A more recent study of this problem revealed little improvement in the situation.9 Although the FDA has called for better alignment between the demographics of study patients and the expected users of a new product, it has not been effective in making this a requirement. The mismatch is most acute for subjects aged >75 years and for patients of all ages with important comorbidities.

A third and relatively inexpensive means of improving the detection of adverse effects is the study design itself. Whether conducted before or after approval, a trial designed principally to enhance marketability is less likely to yield useful adverse event information than one designed to explore a drug’s benefit-risk characteristics evenly-handed. The rofecoxib example cited by Roth-Cline offers a veritable museum of such problematic strategies. Despite evidence that selective cyclooxygenase-2 inhibition might be prothrombotic compared with traditional nonsteroidal antiinflammatory drugs, the Vioxx Gastrointestinal Outcomes Research (VIGOR) trial prohibited use of cardioprotective doses of aspirin, even in a population of rheumatoid arthritis patients at increased risk of cardiac disease.6 Similarly, the Adenomatous Polypl Prevention on Vioxx (APPROVe) trial, which ultimately led to the drug’s withdrawal, enrolled a relatively young cohort and excluded patients with recent ischemic cardiovascular disease.10 Avoiding blatant risk-obscuring aspects of trial design is a better way to elucidate safety problems than adjusting the probability value required for rejection of the null hypothesis for efficacy. There is also much promise in better use of pharmacogenetics, biomarkers, and other basic science approaches to provide earlier warnings about drug toxicities, a strategy the FDA has begun to embrace.11

Recent evidence published in the Federal Register might also seem to lend credence to the tempting idea that extending preapproval studies is the only sure means we have of identifying adverse events. Each year, the FDA reports to Congress on the progress made by pharmaceutical companies in conducting postmarketing studies mandated by the agency. In what has become an annual pattern, the FDA once again reported that of 1231 active “postmarketing commitments,” many of which had been mandated as a condition of approval, nearly two thirds had not even been started.5

How could this be? After the FDA has approved a product, it has few regulatory tools available to compel companies to conduct follow-up studies. Once approval is granted, its regulatory authority is limited primarily to the “nuclear option” of taking the drug off the market. Ironically, even that remedy can be delayed if the safety studies needed to document a serious risk were never performed.

This dismal record makes it easier to understand why some might favor forcing companies to measure risks more thoroughly during the only period when the government has any real influence: before approval. In the case of rofecoxib, however, this was not the issue. There were already ample signals in preapproval studies of possibly increased cardiovascular risk and even a documented (and statistically significant) 5-fold increase in myocardial infarction rates in a trial published soon after marketing began.6 The problem was not that these risks were not detected early enough; it was that they were not acted on appropriately, by the manufacturer or by the FDA, once they were noted. A more sensible policy would be to ensure that needed safety studies are required as promptly as possible, either as part of the approval process or immediately thereafter. These might be rigorous observational studies of drug use and adverse events in large, well-defined populations, or new clinical trials that target specific clinical questions. This could be accomplished if the FDA were able to compel companies to conduct such research or if there were other funding streams available to support these studies. Holding a new drug’s approval hostage to excessively stringent statistical requirements is an inefficient means of accomplishing this important goal.

Wrong Lesson 2: Either We Can Approve Drugs in a Timely Manner or We Can Learn Enough About Their Risks Before Approval, but We Cannot Do Both

This flawed assumption might be called the “Heisenberg fallacy.” Such a measurement paradox may apply to electrons but not to drugs. Defenders of this view argue that improving the capacity of preapproval studies to detect risks will inevitably delay the availability of new products and drive up their costs. In using this argument to reject the extreme efficacy evidence proposal, Roth-Cline arrives at the right conclusion for the wrong reason. Drug evaluation is not a zero-sum game requiring impossible tradeoffs among numbers of subjects, probability values, and time, as Roth-Cline suggests. What matters far more is that studies be designed primarily to generate scientific knowledge, not just to grease the wheels toward rapid approval or expansion of a product’s uses. At present, the design of many preapproval trials is more of a problem than the probability values used to evaluate them.

In response to decades of pressure to make the FDA more industry-friendly, we have drifted toward a lowest-possible-standard approach. Often, a manufacturer must show merely that its new product works better than a comparator (often placebo) in achieving a surrogate outcome (eg, improvement in a laboratory test) in a modestly sized sample of atypically younger and healthier patients (compared with expected users of the marketed drug) who are observed over a brief period of time (weeks or a few months, even for medications designed to be used chronically). There are several ways these standards could be made more relevant to practice without dramatically increasing the duration or cost of drug evaluation and without recalibrating the level of statistical certainty required for approval.7

For toxicities that do not depend on the duration of drug exposure, risk detection can be enhanced without extending trial length simply by enrolling more adequate numbers of patients. An even better way to identify potential safety problems would be to correct the present maldistribution of age and comorbidity among study subjects. More than a decade ago, our group reported on the systematic exclusion of older patients in pivotal drug trials in cardiovascular disease.8 A more recent study of this problem revealed little improvement in the situation.9 Although the FDA has called for better alignment between the demographics of study patients and the expected users of a new product, it has not been effective in making this a requirement. The mismatch is most acute for subjects aged >75 years and for patients of all ages with important comorbidities.

A third and relatively inexpensive means of improving the detection of adverse effects is the study design itself. Whether conducted before or after approval, a trial designed principally to enhance marketability is less likely to yield useful adverse event information than one designed to explore a drug’s benefit-risk characteristics evenly-handed. The rofecoxib example cited by Roth-Cline offers a veritable museum of such problematic strategies. Despite evidence that selective cyclooxygenase-2 inhibition might be prothrombotic compared with traditional nonsteroidal antiinflammatory drugs, the Vioxx Gastrointestinal Outcomes Research (VIGOR) trial prohibited use of cardioprotective doses of aspirin, even in a population of rheumatoid arthritis patients at increased risk of cardiac disease.6 Similarly, the Adenomatous Polypl Prevention on Vioxx (APPROVe) trial, which ultimately led to the drug’s withdrawal, enrolled a relatively young cohort and excluded patients with recent ischemic cardiovascular disease.10 Avoiding blatant risk-obscuring aspects of trial design is a better way to elucidate safety problems than adjusting the probability value required for rejection of the null hypothesis for efficacy. There is also much promise in better use of pharmacogenetics, biomarkers, and other basic science approaches to provide earlier warnings about drug toxicities, a strategy the FDA has begun to embrace.11

Wrong Lesson 2: Either We Can Approve Drugs in a Timely Manner or We Can Learn Enough About Their Risks Before Approval, but We Cannot Do Both

This flawed assumption might be called the “Heisenberg fallacy.” Such a measurement paradox may apply to electrons but not to drugs. Defenders of this view argue that improving the capacity of preapproval studies to detect risks will inevitably delay the availability of new products and drive up their costs. In using this argument to reject the extreme efficacy evidence proposal, Roth-Cline arrives at the right conclusion for the wrong reason. Drug evaluation is not a zero-sum game requiring impossible tradeoffs among numbers of subjects, probability values, and time, as Roth-Cline suggests. What matters far more is that studies be designed primarily to generate scientific knowledge, not just to grease the wheels toward rapid approval or expansion of a product’s uses. At present, the design of many preapproval trials is more of a problem than the probability values used to evaluate them.

In response to decades of pressure to make the FDA more industry-friendly, we have drifted toward a lowest-possible-standard approach. Often, a manufacturer must show merely that its new product works better than a comparator (often placebo) in achieving a surrogate outcome (eg, improvement in a laboratory test) in a modestly sized sample of atypically younger and healthier patients (compared with expected users of the marketed drug) who are observed over a brief period of time (weeks or a few months, even for medications designed to be used chronically). There are several ways these standards could be made more relevant to practice without dramatically increasing the duration or cost of drug evaluation and without recalibrating the level of statistical certainty required for approval.7

For toxicities that do not depend on the duration of drug exposure, risk detection can be enhanced without extending trial length simply by enrolling more adequate numbers of patients. An even better way to identify potential safety problems would be to correct the present maldistribution of age and comorbidity among study subjects. More than a decade ago, our group reported on the systematic exclusion of older patients in pivotal drug trials in cardiovascular disease.8 A more recent study of this problem revealed little improvement in the situation.9 Although the FDA has called for better alignment between the demographics of study patients and the expected users of a new product, it has not been effective in making this a requirement. The mismatch is most acute for subjects aged >75 years and for patients of all ages with important comorbidities.

A third and relatively inexpensive means of improving the detection of adverse effects is the study design itself. Whether conducted before or after approval, a trial designed principally to enhance marketability is less likely to yield useful adverse event information than one designed to explore a drug’s benefit-risk characteristics evenly-handed. The rofecoxib example cited by Roth-Cline offers a veritable museum of such problematic strategies. Despite evidence that selective cyclooxygenase-2 inhibition might be prothrombotic compared with traditional nonsteroidal antiinflammatory drugs, the Vioxx Gastrointestinal Outcomes Research (VIGOR) trial prohibited use of cardioprotective doses of aspirin, even in a population of rheumatoid arthritis patients at increased risk of cardiac disease.6 Similarly, the Adenomatous Polypl Prevention on Vioxx (APPROVe) trial, which ultimately led to the drug’s withdrawal, enrolled a relatively young cohort and excluded patients with recent ischemic cardiovascular disease.10 Avoiding blatant risk-obscuring aspects of trial design is a better way to elucidate safety problems than adjusting the probability value required for rejection of the null hypothesis for efficacy. There is also much promise in better use of pharmacogenetics, biomarkers, and other basic science approaches to provide earlier warnings about drug toxicities, a strategy the FDA has begun to embrace.11
Wrong Lesson 3: Rofecoxib Proved That Many Important Adverse Events Are Inherently Unknowable Until It Is Too Late, and Therefore We Are Unlikely to Develop Better Systems for Their Detection

Ample data are available to refute this nihilistic view, with numerous examples from the cardiovascular literature alone. Evidence of the potential thrombotic effects of cyclooxygenase-2 inhibitors was present even in the early stages of development of these drugs but was not used to develop trials designed to elucidate this problem. Similarly, data on the risk of pulmonary hypertension caused by the diet aid fenfluramine (and its d-isomer, dexfenfluramine [Redux, Wyeth-Ayerst]) were in place long before those drugs were taken off the market in 1997. Concerns about potential pharmacokinetic problems with the calcium-channel blocker mibebradil (Posicor, Roche) had been recorded well before its withdrawal in 1998. Finally, cerivastatin (Baycol, Bayer) was known to pose a greater risk of substantial creatine kinase elevation than other statins (predictive of the rhabdomyolysis that followed) even before the drug was approved. For each, the problem was not inadequate statistical stringency as the criterion for approval, although ironically such a requirement would have increased the chances that these already-apparent problems would be noticed before the use of these drugs became widespread. Rather, zealous promotion was combined with failure of regulatory insight or will, followed by clinicians’ overly enthusiastic adoption of flawed but aggressively marketed products.

Wrong Lesson 4: Greater Scrutiny for Adverse Events Will Impede the Availability of Important New Products and Drive Up Drug Prices

The main reason to avoid using the “probability value lever” as a primary tool of drug risk assessment is that it is a blunt instrument, not because it will dramatically slow the development of needed drugs. Most of the withdrawn drugs in Roth-Cline’s Table 1 were treatments for conditions for which satisfactory therapies already existed and were not striking clinical breakthroughs. In those cases, a slower approval time would not have had any clinical disadvantage and, we now know, would have indeed protected the public’s health. The same applied to rofecoxib, which was never a more powerful analgesic than older nonsteroidal antiinflammatory drugs. The modest gastroprotection afforded by selective cyclooxygenase-2 inhibitors is probably about the same as that afforded by a proton pump inhibitor and is largely negated by concurrent use of low-dose aspirin.

As for cost, the United States almost certainly spends more on medications because of the current flawed system than it would with a better approach to drug evaluation. The rofecoxib example is apt here as well: Although the drug provided no efficacy advantage and greater all-cause risk, it cost Americans $2.5 billion a year before its withdrawal. Much of that was public money; an earlier workup of the numerous signals of this one drug’s safety problems would have saved far more than it would cost to put a better system in place nationally.

Wrong Lesson 5: The Best Strategy Now Is for the FDA to Warn About Everything

The public-communications correlate of using an unreasonably high threshold for efficacy is using an unreasonably low threshold for warning about possible risks. This may account for the FDA’s hyperactive behavior after its embarrassment over not adequately addressing the risks of rofecoxib until the drug was withdrawn. Several months later, the agency recommended including a “black box warning” about cardiac risk in the official labeling of all nonsteroidal antiinflammatory drugs, regardless of whether there was good evidence of a safety problem or not. Similar calls were made recently for warnings about the cardiac risks of attention deficit disorder drugs, again without compelling data. (Evidence of the durability of this problem is the fact that such evidence was never assembled even though one member of the class, methylphenidate [Ritalin, Novartis], has been in use for about 50 years.) Obtundation and spasticity are both suboptimal states of arousal. In the absence of an adequate approach to collecting and analyzing data, however, it is understandable that an evidence-deprived agency might resort to either kind of extreme reaction.

Evaluation of drug safety has much in common with evaluation of a patient, in that both are inherently Bayesian processes. Armed with an informed set of prior probabilities, one looks for signals. Suggestive pieces of evidence are then worked up further, even if they do not initially offer black-and-white confirmation of “significance.” Additional targeted studies are then conducted in a timely way to follow up on promising hypotheses. This is not a process that can be adequately crammed into the first evaluation of a drug any more than a thoughtful clinical workup can be completed in the first moments of a hospitalization or office visit. Because this comprehensive assessment cannot be done “once and for all” at the time of initial evaluation, it has been proposed that the drug approval process be divided into 2 steps: an initial clearance for marketing after a rapid but rigorous review, followed by a reappraisal 2 or 3 years later that takes account of subsequent experience with both safety and effectiveness. That would fit the real nature of drug benefit-risk assessment far better than simply slowing down inappropriately designed and minimally generalizable preapproval trials.

It is a sad commentary on the current state of drug evaluation and safety surveillance in the United States that some believe that the only way we can be sure of getting adequate safety information is by requiring extreme criteria for demonstrating efficacy. Fortunately, other solutions are possible that are more scientifically rigorous, clinically appropriate, and feasible. All that is lacking is the political will to implement them.

Disclosures

Dr Avorn reports having received research grants from Pfizer and Merck to study adverse effects of cyclooxygenase-2 inhibitors. He
has served as a consultant with attorneys in the Vioxx-related litigation but received no remuneration for this work.

References
Evaluating Drug Effects in the Post-Vioxx World: There Must Be a Better Way
Jerry Avorn

Circulation. 2006;113:2173-2176
doi: 10.1161/CIRCULATIONAHA.106.625749
Circulation is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2006 American Heart Association, Inc. All rights reserved.
Print ISSN: 0009-7322. Online ISSN: 1524-4539

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circ.ahajournals.org/content/113/18/2173

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Circulation is online at:
http://circ.ahajournals.org/subscriptions/