2

Drug Regulation and the Inductive Risk Calculus

Jacob Stegenga

Introduction

Drug regulation is fraught with inductive risk. Regulators must make a prediction about whether or not an experimental pharmaceutical will be effective and relatively safe when used by typical patients, and such predictions are based on a complex, indeterminate, and incomplete evidential basis. Such inductive risk has important practical consequences. If regulators reject an experimental drug when it in fact has a favorable benefit/harm profile, then a valuable intervention is denied to the public and a company's material interests are needlessly thwarted. Conversely, if regulators approve an experimental drug when it in fact has an unfavorable benefit/harm profile, then resources are wasted, people are needlessly harmed, and other potentially more effective treatments are underutilized. Given that such regulatory decisions have these practical consequences, non-epistemic values about the relative importance of these consequences impact the way such regulatory decisions are made (similar to the analysis of laboratory studies on the toxic effects of dioxins presented in Douglas [2000]). To balance the competing demands of the pertinent non-epistemic values, regulators must perform what I call an "inductive risk calculus."

At least in the American context this inductive risk calculus is not well-managed. The epistemic standard with which the US Food and Drug Administration (FDA) assesses the effectiveness and harm profile of experimental drugs is low. That is, the evidence that the FDA requires for assessing the safety and effectiveness of new pharmaceuticals is insufficient to make a reliable inference about the safety and effectiveness of new pharmaceuticals. The usual FDA requirement for a drug to be approved for general use is two "phase 3" randomized controlled trials in which the experimental drug is deemed more efficacious than placebo (or other comparator control substances). There are a number of problems with this standard. The standard does not take into account the number of trials which have been carried out, and given the ubiquitous phenomenon of publication bias, two positive clinical trials of an experimental drug does not warrant a conclusion that the drug is truly efficacious. Even if the drug is truly efficacious in the experimental context, there are many reasons why the drug might not be effective in a general context. Moreover, this epistemic standard is not a reliable guide to assessing the harm profile of experimental drugs, for a number of subtle reasons. I detail these and other problems for the epistemic standard of drug approval. In short, I show that even if, in some particular case, the explicit epistemic standard is met, there are a variety of more subtle factors that can render the available evidence dramatically unreliable.

The inductive risk calculus for drug approval would be better managed if the epistemic standard for drug approval were enhanced. I argue that the epistemic standard for drug approval in the United States should be enhanced in a variety of ways. This, though, increases the practical risk that regulators might reject more experimental drugs that in fact have favorable benefit/harm profiles, thereby denying valuable interventions to the public and thwarting commercial interests. How worrying is this consequence of raising the epistemic standards for drug approval? I argue: not very. There simply have not been many very effective drugs introduced into the pharmaceutical arsenal in recent generations, and besides, effective drugs would still be approved if epistemic standards for assessing experimental drugs were enhanced.

I illustrate these arguments with a number of examples. A running example is based on the drug rosiglitazone (trade name Avandia), which was recently the world's best-selling drug for type-2 diabetes. The evidence surrounding the safety and efficacy of rosiglitazone was shrouded in secrecy, thereby illustrating the problem of publication bias. A major trial testing the safety of rosiglitazone involved screening the research subjects with a large number of inclusion and exclusion criteria (a typical practice), thereby illustrating the insensitivity of the FDA standard to the problem of extrapolation from controlled research settings to real-world clinical settings. Rosiglitazone ended up being more harmful than was thought at the time of FDA approval, thereby illustrating the insufficient attention to the harm profile of experimental drugs in the FDA standard.

Non-epistemic values influence one's stance on an inductive risk calculus, especially in empirical contexts in which evidence informs policy—this is the conclusion of the argument from inductive risk (see, e.g., Douglas 2000; Elliott and McKaughan 2009). In some cases the particular influence of non-epistemic values on an inductive risk calculus is warranted while in other cases the influence is pernicious. Thus, we must demarcate the former from the latter—some stances on an inductive risk calculus are justified while others are not. Torsten Wilholt (2009) notes that a general and principled criterion for such demarcation has proven to be elusive (Wilholt offers such a criterion himself, but I argue that this criterion is neither necessary nor sufficient to demarcate pernicious from permitted influences of non-epistemic values on an inductive risk calculus). One might despair—without such a demarcation criterion we lose touch with objectivity. Corporate scientists who tweak every detail of experimental design in such a way that shareholder profit is maximized are just as objective—goes this despair—as regulatory epidemiologists who interpret the evidence from those experiments with sole regard to protecting the health of the public. However, in this chapter I show that, at least within a particular domain, rational deliberation about one's stance on an inductive risk calculus is possible even in the absence of a general principle regarding the influence of non-epistemic values on an inductive risk calculus.

Drug Approval in the United States

The Center for Drug Evaluation and Research (CDER) is a branch of the FDA that is responsible for regulating new drug approval. If a company wants to introduce a new pharmaceutical into the US market, it must submit a "new drug application" to CDER. The primary role of CDER is to evaluate the new drug application to determine if the new drug is (to use the FDA phrase) "safe and effective when used as directed."

There are multiple steps leading up to a new drug application. To begin, the institutions responsible for the experimental pharmaceutical (the "sponsors," including pharmaceutical companies, universities, and other research organizations) must test the experimental pharmaceutical in laboratory animals. If the results of animal tests are promising enough, the sponsors submit what is called an "investigational new drug application" to the FDA to get approval to begin human clinical trials. Initial tests in humans are performed in "phase 1" trials, which usually have less than one hundred healthy volunteers, and are intended to discover the most important harmful effects of the drug. If the drug appears to be not excessively toxic in a phase 1 trial, then "phase 2" trials might be initiated. Phase 2 trials are randomized controlled trials (RCTs) which usually involve a couple of hundred subjects, and are intended to gather more data on harms caused by the pharmaceutical while also testing the efficacy of the pharmaceutical in patients with the disease meant to be treated. If the drug appears to have some efficacy in phase 2 trials, "phase 3" trials are performed. Phase 3 trials are also RCTs which usually have several hundred to several thousand subjects, and are intended to gather more precise data on the efficacy of the experimental drug. It usually takes around ten years to go from pre-clinical animal studies to the completion of phase 3 trials. The FDA does not conduct its own studies; it relies on the data submitted by the sponsor of the new drug application.

If a sponsor deems the drug promising enough, they submit a "new drug application" to the FDA. The FDA puts together a review team to assess the new drug application; the review team usually includes physicians, statisticians, pharmacologists, and other scientists. The principal question addressed by the review team is whether or not the new drug is safe and effective. If the new drug application is approved, then the drug may be sold to consumers. At this point, the FDA may require the sponsors to conduct "phase 4" studies, which are trials or observational studies used for assessing safety and effectiveness of the drug after the drug has been approved for general public use.

The epistemic standard for meeting the "safe and effective" requirement is ultimately decided on a case-by-case basis depending on various contextual factors. However, there are some common elements of the epistemic standard. The evidence submitted by a sponsor must include an RCT in which the results are deemed "positive." A positive trial, according to the FDA, is one in which an experimental group in the trial appears to gain some benefit from the experimental intervention compared to the control group (which in typical cases receives either placebo or a competitor drug), and this apparent benefit is deemed "statistically significant" in that the *p* value of a frequentist statistical test on this result is less than .05. In other words, a positive trial is one in which there is less than a 5% probability that one would observe such a difference in the measured parameter between the trial's intervention group and control group if the "null" hypothesis were true (the null hypothesis is usually the hypothesis that the intervention is not effective). The FDA has generally required two positive trials to establish effectiveness and thereby approve the new drug application (CDER 1998). The FDA sometimes makes exceptions to the two-positive-trial rule, approving a new drug application on the basis of a single trial which might be supplemented with other confirming evidence, such as evidence from related positive trials or animal studies, and sometimes does away with required supplemental evidence if the RCT happens to be a large multi-center trial. The measured parameter in an acceptable trial can be an important patient-level outcome (such as death), but the FDA also accepts trials which only measure "surrogate endpoints," which are "laboratory measures or other tests that have no direct or obvious relationship to how a patient feels or to any clinical symptom, but on which a beneficial effect of a drug is presumed to predict a desired beneficial effect on such a clinical outcome" (Katz 2004, 309). In short: to approve a new drug, generally the FDA requires two RCTs in which the drug appears to have a statistically significant benefit. I will articulate problems with this standard, but first I introduce the notion of an "inductive risk calculus."

The Inductive Risk Calculus

Some critics argue that the FDA overregulates the introduction of new pharmaceuticals. These critics hold that the epistemic standards required for new drug approval are cumbersome, disincentivize research into new pharmaceuticals, and raise the prices of drugs. Such criticisms tend to come from free-market economists or institutions (see, e.g., Becker 2002; Friedman and Friedman 1990). Other critics argue that the FDA underregulates the introduction of new pharmaceuticals. These critics hold that the epistemic standards required for new drug approval are too low and allow drugs that are relatively ineffective or unsafe to be approved. Such criticisms have been voiced by academic scientists (such as Steve Nissen, who performed the 2007 meta-analysis on rosiglitazone), scientific organizations (such as the US Institute of Medicine), and even by staff within the FDA (such as the epidemiologist David Graham) (see, e.g., Carozza 2005; Institute of Medicine 2006).

Just as in the prominent discussion of inductive risk presented in Richard Rudner (1953) and extended by Heather Douglas (2000) and others, non-epistemic values play a role in setting epistemic standards in policy contexts. When assessing the effectiveness and safety of a pharmaceutical, one is liable to make a false inference based on the available evidence—accordingly, one faces inductive risk. At least some experimental pharmaceuticals are effective (though many are not), and few experimental pharmaceuticals are completely safe, since most cause at least some unintended harmful effects. Regulators must make a judgment about the relative effectiveness-harm profile of an experimental pharmaceutical, based on whatever evidence they have. To do this, regulators must make an inference, and there are two fundamental errors they can make in this context: they can approve a drug as having a favorable effectiveness-harm profile when it in fact does not, or they can reject a drug as not having a favorable effectiveness-harm profile when it in fact does. The former kind of error (unwarranted drug approvals) can harm patients by allowing relatively ineffective or unsafe drugs to be available, and the latter kind of error (unwarranted drug rejections) can harm patients by prohibiting relatively effective or safe drugs from being available and can harm the financial interests of the manufacturer of the drug.

To avoid these two fundamental kinds of errors, regulators employ numerous tactics. Many of these tactics tradeoff against each other, in that employing a tactic to decrease the probability of committing one of the error types increases the probability of committing the other error type. For example, demanding more positive RCTs for drug approval decreases the probability of unwarranted drug approvals but increases the probability of unwarranted drug rejections. Or to take an extreme case, a tactic to guarantee that regulators never commit the error of unwarranted drug rejections is to approve all new drug applications, thereby greatly increasing the probability of unwarranted drug approvals; and vice versa, a tactic to guarantee that regulators never commit the error of unwarranted drug approvals is to reject all new drug applications, thereby greatly increasing the probability of unwarranted drug rejections. Thus, we can conceptualize a scale of inductive risk: on one end of the scale is certainty that the error of unwarranted approvals is avoided (and thus a high probability that the error of unwarranted drug rejections is committed) and on the other end of the scale is certainty that the error of unwarranted drug rejections is avoided (and thus a high probability that the error of unwarranted drug approvals is committed). Between these two extreme ends of the scale of inductive risk are intermediate positions.

Regulators must determine where their policies stand on this scale of inductive risk. This is an *inductive risk calculus*. Non-epistemic values influence this inductive risk calculus (Douglas 2009; Elliott 2011). The criticisms of FDA overregulation or underregulation can be understood in terms of this calculus: some critics hold that the FDA's inductive risk calculus places its regulatory stance too far toward the extreme of never committing the error of unwarranted drug approvals (overregulation), whereas other critics hold that the FDA's inductive risk calculus places its regulatory stance too far toward the other extreme of never committing the error of unwarranted drug rejections (underregulation). In the next section, I argue that there are numerous problems with the FDA epistemic standards for new drug applications; these considerations lend support to those who challenge the FDA with underregulation. In the section following, I suggest some ways in which the inductive risk calculus can be retuned to address some of these problems. In the final section, I argue that the principal arguments of those who challenge the FDA with overregulation are not compelling.

Problems with the Food and Drug Administration Standard

There are numerous problems with the FDA epistemic standard for drug approval; these problems amount to the epistemic standard for drug approval being too low. Although the epistemic requirements for drug approval described above sound cumbersome, in the context of contemporary biomedical research they are too easy to satisfy with respect to any reasonable norm of evaluation. Consider Philip Kitcher's notion of well-ordered certification applied to the inductive risk calculus (Kitcher 2011): certification is well-ordered just in case ideal deliberation would endorse the certifier's stance on an inductive risk calculus. The FDA is involved in certification when they assess new drug applications. Ideal deliberators would conclude that the inductive risk calculus of the FDA stands too far toward the extreme of never committing the error of unwarranted drug rejections—in other words, the FDA underregulates. That is the argument of this section.

A fundamental problem is that the FDA does not conduct its own studies of the drugs under question, nor does it examine other data that might be available on the drugs from other organizations (including academic, industrial, or government organizations). Although the trials that industrial sponsors must perform to support a new drug application are constrained by structural standards for trial design (for example, trials must be randomized), there is still a wide degree of latitude in how studies are designed, executed, and analyzed, and this permits biases to enter the research. Since manufacturers of pharmaceuticals have a very strong financial incentive to demonstrate effectiveness of their products, they may exploit this researcher latitude in such a way that their products appear to be more effective and less harmful than they truly are (I argue this point in more detail in Stegenga [forthcoming]).

A more concrete problem with the FDA standard for drug approval is that a standard based on statistical significance lends itself to "p-hacking." Spurious correlations can occur by chance, and the more complex a data set is, and the more analyses performed on a data set, the more likely it is that one will discover a spurious correlation. P-hacking can occur when a researcher exercises "researcher degree of freedom": researchers perform multiple studies, on multiple parameters, choosing which parameters to measure and which comparisons to make and which analyses to perform, and they can do this until they find a low enough *p* value to satisfy the standard of statistical significance. Since low *p* values are likely to occur by chance alone, p-hacking makes it easy to satisfy the standard of statistical significance even when the experimental drug is not in fact beneficial. P-hacking can be mitigated if trial designs explicitly state, in advance, what primary outcomes will be measured and how the data will be analyzed. Unfortunately, a recent study found that, for trials with pre-designated clinical trial plans, about half of clinical trials had at least one primary outcome that was changed, introduced, or omitted (Dwan et al. 2008).

Even when no p-hacking occurs, a statistically significant result in a trial does not entail that a clinically significant result has been found. This is for a number of reasons. The result, although statistically significant, may be due to chance. The result, although statistically significant, may be clinically meaningless because the effect size is tiny. The result, although statistically significant, may be clinically meaningless because the subjects in the trial differed in important ways from typical patients.

This latter issue is widespread. Trials employ a number of exclusion and inclusion criteria when recruiting subjects for a trial, which has the effect of rendering study populations very different from typical patients. Inclusion criteria stipulate necessary features that patients must have to be included in a trial, and exclusion criteria stipulate features that patients must necessarily not have else they are excluded from a trial. Typical patients tend to be older, on more drugs, and have more diseases than trial subjects, and these differences are known to modulate the effectiveness and harmfulness of pharmaceuticals. A major trial testing rosiglitazone provides a good example of this: the RECORD trial employed seven inclusion criteria and sixteen exclusion criteria, and a result of these criteria was that subjects in the trial were, on average, healthier than typical patients; for example, subjects in the trial had a heart attack rate about 40% less than that of the equivalent demographic group (middle-aged people with type-2 diabetes) in the broader population.

Another problem with the FDA standard for drug approval is that although the effect size of a trial might be statistically significant, the measured parameter in the trial might be clinically irrelevant. For an example of this problem, consider clinical trials on antidepressants. These trials employ a measurement tool called the Hamilton Rating Scale for Depression (HAMD). This scale has a number of questions which are scored and summed, and the overall score, with a maximum of about fifty, is said to be a measure of the intensity of one's depression. The best assessments of antidepressants conclude that antidepressants on average lower HAMD scores by less than three points (Kirsch et al. 2008). However, the HAMD scale includes up to six points on quality of sleep and four points on the extent to which one fidgets. Thus, a drug's capacity to decrease one's HAMD score by three points does not indicate that the drug will be helpful in mitigating core symptoms of depression, because it might simply modulate fidgeting or cause slight improvements to sleep (for more on problems of measurement in clinical research, see Stegenga 2015). The FDA standard is too permissive regarding which parameters must be measured and modified by an experimental drug in a clinical trial.

Putting aside all of the problems with the "statistical significance" standard, there is a more technical and fundamental problem with this standard. To articulate this problem will require a brief use of formalisms. Suppose: our hypothesis of interest (H) is that a drug is effective, the null hypothesis (H_o) is that the drug is not effective, and a trial generates evidence (E) that suggests that the drug is effective with a p value of .05. The FDA standard, which is satisfied in this case, is based on the probability that we would get E if H_o were true: $P(E|H_o)$. But the FDA must determine if the drug is effective: the FDA must estimate how probable H is now that we have E: $P(H|E)$. There is a very widespread habit of assuming that one can directly infer $P(H|E)$ from $P(E|H_o)$. But this is fallacious—such inferences commit what is called the base-rate fallacy. To see this, apply Bayes's Theorem to $P(H|E)$:

$$P(H|E) = P(E|H)P(H) / \left[P(E|H)P(H) + P(E|H_0)P(H_0)\right].$$

The statistical significance level, or p value, only indicates $P(E|H_o)$, which, as one can see by examining the equation, is grossly insufficient to infer $P(H|E)$ (because, to infer $P(H|E)$, in addition to taking into account $P(E|H_o)$, one also needs to take into account $P(H)$ and $P(H_o)$). Yet the p value is the epistemological basis of the FDA standard. Thus, the epistemological basis of the FDA standard is grossly insufficient for the inference it is required to make.

Consider a radical example of a study with a low p value in which the absurdity of the base-rate fallacy is obvious. A researcher tested the efficacy of remote, retroactive intercessory prayer for patients with bloodstream infections, and found that length of hospital stay and duration of fever was shorter among patients who were retroactively prayed for compared with control patients, and these findings had p values of less than .05 (Leibovici 2001).

Yet it would be absurd to conclude that this evidence justifies belief in remote retroactive intercessory prayer—in other words, it would be absurd to conclude that $P(H|E)$ is high. That is because $P(H)$ ought to be very low—our prior expectation that remote retroactive intercessory prayer is effective ought to be very low. As the equation indicates, $P(H|E)$ is directly proportional to $P(H)$, and so a low $P(H)$ will render $P(H|E)$ lower than it otherwise would have been had $P(H)$ been higher. Inferring effectiveness of remote retroactive intercessory prayer on the basis of the low p value in this study would be fallacious.

Trials are often too short in duration and too small in number of subjects to detect rare harms of drugs or harms that take months or years to manifest (Stegenga 2016). Moreover, short-duration trials might be able to detect short-term benefits of the drug despite an absence of long-term benefits. For example, recent meta-analyses have shown that corticosteroid injections for knee arthritis decrease patients' pain for about a week, but have no benefit in the longer run; however, because corticosteroid injections for knee arthritis were studied with short-term trials for many years, they were wrongly thought to provide substantial and lasting benefits to patients with knee arthritis. The FDA standard does not account for the short duration of most trials.

Perhaps the most worrying problem about the FDA standard is that it does not take into account publication bias, in which positive trials are published but negative trials remain unpublished. The two-positive-trials rule can be satisfied by a new drug application even if many trials generated evidence that suggested that the drug is not effective—as long as there are two positive trials, the standard is satisfied. To illustrate publication bias, consider reboxetine. Reboxetine is an antidepressant marketed in Europe. Recently a meta-analysis was performed in which the researchers had access to both published and unpublished data (Eyding et al. 2010). Of the thirteen trials that had been performed on reboxetine, data from 74% of patients remained unpublished. Seven of the trials compared reboxetine against placebo: one had positive results and only this one was published; the other six trials (comprising almost ten times as many patients) gave null results, and none of these were published. The trials that compared reboxetine to competitor drugs were worse. Three small trials suggested that reboxetine was superior to its competitors. But the other trials, with three times as many patients, showed that reboxetine was less effective than its competitors and had worse side effects (for a discussion of this case, see Goldacre 2012).

Publication bias can also mask the harms of new drugs. One study estimated the publication rate of phase 1 trials at less than 10% (Decullier, Chan, and Chapuis 2009), which is extremely concerning given that phase 1 trials are the foundation for assessing the harm profile of drugs generally. Of course, publication bias also affects phase 3 RCTs.

The drug rosiglitazone provides a striking illustration of publication bias of phase 3 trials. In this case, the FDA itself contributed to the secrecy associated with publication bias. Steve Nissen, an expert in type-2 diabetes, requested data from GlaxoSmithKline (GSK), the manufacturer of rosiglitazone, but GSK refused to share the data. However, the company had earlier been required to develop a registry of their clinical trial data (as the result of a legal settlement for fraud pertaining to its drug paroxetine, or Paxil). Nissen identified forty-two RCTs of rosiglitazone, but only seven of these trials had been published. Nissen performed a meta-analysis on all of the trials, and his analysis concluded that rosiglitazone increases the risk of cardiovascular harms by 43%. Nissen submitted his meta-analysis to the *New England Journal of Medicine*, and one of the peer reviewers faxed a copy to GSK. In an internal email the director of research at the company subsequently wrote "FDA, Nissen, and GSK all come to comparable conclusions regarding increased risk for ischemic events, ranging from 30% to 43%!" In short, the FDA and GSK already knew of the cardiovascular harm caused by rosiglitazone, but neither organization had publicized this finding.

A survey of FDA reviewers indicated that even those involved in the drug approval process believe that the epistemic standards are too low—many FDA reviewers expressed concern about the low standards for evaluating effectiveness and harmfulness of drugs (Lurie and Wolfe 1998). One reviewer claimed that the FDA leans toward approving "everything." Reviewers even reported cases in which they recommended that new drug applications be rejected and the drugs were nevertheless approved. In another context, a well-known epidemiologist and associate director of the FDA's Office of Pharmacovigilance and Epidemiology (formerly Office of Drug Safety) claimed that the "FDA consistently overrated the benefits of the drugs it approved and rejected, downplayed, or ignored the safety problems . . . when FDA approves a drug, it usually has no evidence that the drug will provide a meaningful benefit to patients" (Carozza 2005, 39–40). Thus far in the year of writing this chapter (September 2015), the FDA's new drug application approval rate is 88% when taking into account multiple new uses of a new drug; if one takes into account solely the number of drugs under consideration, the FDA has rejected one drug and approved twenty-three, for an approval rate of 96%.

The problems described in this section entail that the current FDA standard for new drug approval is low. In other words, the FDA's inductive risk calculus for new drug approval lies far toward the extreme of avoiding unwarranted drug rejections. The epistemic standard should be raised to achieve a more balanced inductive risk calculus.

Retuning the Inductive Risk Calculus

A general way to justify a particular stance on an inductive risk calculus would be to appeal to a principled criterion that excludes stances which are constituted by unwarranted influence of non-epistemic values. What might such a criterion look like? What renders the influence of some non-epistemic values justified and others unjustified? Wilholt (2009) argues that the influence of non-epistemic values on an inductive risk calculus is impermissible when it involves infringement of the conventional standards held by the pertinent research community. A problem with this principle is that we have already seen that the conventional standard that is explicitly articulated in the domain of pharmaceutical regulation—the two-positive-trials standard—is far too easy to satisfy and can be satisfied in cases in which the evidence is unreliable with respect to the safety and effectiveness of experimental pharmaceuticals. Thus, infringement of the conventional standards held by the pertinent research community is unnecessary for a stance on an inductive risk calculus to be unjustified. Moreover, in some cases, infringement of the conventional standards might be justified on epistemic grounds (say, by relaxing the two-positive-trials standard in cases in which there are other grounds for thinking that the experimental pharmaceutical is effective) or non-epistemic grounds (say, for cases in which the experimental pharmaceutical is the last hope for mortally ill patients). Thus, infringement of the conventional standards held by the pertinent research community is insufficient for a stance on an inductive risk calculus to be unjustified.

The FDA's inductive risk calculus should be balanced between the extremes of avoiding unwarranted drug rejections (underregulation) and avoiding unwarranted drug approvals (overregulation). But without a general and principled demarcation criterion, on what grounds can one say that the particular influence of non-epistemic values is justified, or in other words, that one's stance on an inductive risk calculus is warranted? Consider again Kitcher's (2011) notion of well-ordered certification in the context of inductive risk: ideal deliberators pondering an inductive risk calculus—taking into account the relevant non-epistemic values of both patients and manufacturers of pharmaceuticals and society at large—would demand a balanced stance on an inductive risk calculus for drug approval, in which the full range of non-epistemic values is accounted for (in addition, of course, to the full range of epistemic factors). This section provides some guidance for how greater balance could be achieved. Where exactly the FDA's stance should be on the pertinent inductive risk calculus is beside the point—the argument here is that it is currently placed vastly too far toward the position of underregulation and should be significantly shifted toward a more balanced stance.

By appealing to the notion of "balance" in this inductive risk calculus, I do not mean to imply that there is a value-neutral method of determining one's stance on the inductive risk calculus, but rather, that the full range of values should be considered, and that methodological biases should not spuriously shift one's stance on the inductive risk calculus. Earlier, I argued that this is presently not the case. Given the problems with the epistemic standard for drug approval articulated, the fundamental way in which the FDA's inductive risk calculus could achieve more balance is to require more and better evidence regarding the effectiveness and harms of new pharmaceuticals. There are some relatively straightforward tactics to achieve this.

To address the problem of p-hacking, more appropriate quantitative measures of effectiveness should be employed as standards for drug approval. In Stegenga (2015), I argue that effect sizes should be reported using absolute measures such as the "risk difference" measure. The measured effect size should be large enough that a typical patient with the disease in question could expect to receive some substantial benefit from the pharmaceutical on an important patient-level parameter which is pertinent to the disease in question (sadly, as I argue in Stegenga [2015], this is not presently the case). Moreover, trial designs and analytic plans, including the choice of primary outcome to be measured, should be made public in advance of the trial, and departures from the design or analytic plan should mitigate the assessment of the quality of the evidence by the FDA.

Before a new drug application is approved, trials should show that the drug is effective and relatively safe in a broad range of subjects that represents the diversity of typical patients who will eventually use the drug in uncontrolled real-world clinical settings. Trials should be designed to rigorously examine the harm-profile of experimental drugs, and should employ measurement instruments which provide faithful representations of the disease in question.

To address publication bias, all clinical trial data should be made publicly available, and clinical trial registration should be a necessary requirement of all clinical trials for any drug that will eventually be submitted to the FDA for approval (Resnik 2007). The FDA's inductive risk calculus should incorporate all evidence from all trials, and not just two trials that happen to have a positive result. To mitigate the concern about financial conflicts of interest influencing subtle aspects of trial design in a potentially biased manner, the FDA should require evidence from trials performed by organizations which are entirely independent of the manufacturer in question (such as a university or another government agency) (Reiss 2010).

There are structural problems with the way the FDA is organized and funded and how it relates to industry. The FDA epidemiologist David Graham claims that the "FDA is inherently biased in favor of the pharmaceutical industry. It views industry as its client, whose interests it must represent and advance. It views its primary mission as approving as many drugs as it can, regardless of whether the drugs are safe or needed" (Carozza 2005, 39). Much of the funding of CDER comes from user fees paid by industry to have their new drug applications evaluated, and critics claim that since these user fees pay the salaries of reviewers of new drug applications, reviewers are beholden to the sponsors of new drug applications. Moreover, the FDA relies on advisory committees which are composed of internal staff and external scientific consultants, and these committees often have significant conflicts of interest. David Resnik (2007) and Sheldon Krimsky (2003) discuss an investigation which examined 159 meetings by eighteen FDA advisory panels: there was at least one panel member with a financial conflict of interest in 146 of the meetings, and over half the panel members in 88 meetings had financial interests which were "directly related to the topic of the meeting" (Resnik 2007, 25). In other words, most members in most FDA advisory panel meetings had a financial conflict of interest. Finally, critics note that CDER contains both the office that approves new drugs and the office that tracks the harms of drugs that have been approved, which creates an institutional conflict of interest, because once CDER has approved a drug there is a strong disincentive to admit that it made a mistake by paying heed to the office which tracks the harms of approved drugs.

An interesting proposal to address some of the structural problems with the way the FDA is organized and more generally with the imbalanced inductive risk calculus of the FDA is what Justin Biddle (2013) calls "adversarial proceedings for the evaluation of pharmaceuticals." Based on Arthur Kantrowitz's notion of a "science court" (see, e.g., Kantrowitz 1978), this would involve two groups of interlocutors debating the merits of a drug, where one group would be appointed by the sponsor of a drug and the other group would be composed of independent scientists, consumer advocates, and prior critics of the drug. The proceedings would be run by a panel of judges, who would come from a variety of scientific disciplines and would be entirely independent of the drug's sponsor (to Biddle's proposal I would add that philosophers of science—trained in scientific reasoning and knowledgeable about the social context of biomedical research—would be a valuable addition to such panels). Biddle's proposal can be motivated by recent work in feminist epistemology which holds that epistemic standards can be enhanced by including diverse perspectives in scientific evaluation (Wylie 1992). Although the idea would obviously require many details of implementation to be worked out, it is promising and would probably alleviate many of the problems associated with the FDA's imbalanced inductive risk calculus.

Too Radically Retuned?

A counterargument to the view presented here is that increasing the epistemic standards for drug approval will hinder the development of helpful and even life-saving medications, causing people to needlessly suffer. As the eminent economist Gary Becker puts it, "new medicines are a major force behind the rapid advances in both life expectancy and the quality of life that have come during the past 50 years" (2002) and increasing the epistemic standards for drug approval amounts to hindering the development of new drugs, and thus amounts to hindering the great potential of increasing the length and quality of our lives. Even the present nominee for commissioner of the FDA, Dr. Robert Califf, seems to hold a view like this—in a recent presentation Dr. Califf included a slide which claimed that regulation is a barrier to innovation. This is a dubious claim, however, for a number of reasons.

As the historian of medicine Thomas McKeown argued, contrary to the view expressed by Becker, the increase in Western life expectancy has had little to do with medicine and was much more a result of better living standards such as increased nutrition (1976). McKeown's thesis is controversial, but even his critics usually agree that it was factors other than medicine which were responsible for increasing life expectancy, such as sanitary measures and clean drinking water.

Strengthening regulation will not significantly hinder the introduction of novel effective pharmaceuticals. That is because there is in principle a dearth of effective pharmaceuticals, and this dearth is not a result of regulation but rather is a result of the complex nature of diseases and the complex ways in which drugs interact with normal and pathological physiology. Elsewhere I argue that the "magic bullet" model of pharmaceuticals is an ideal standard for drugs. Highly effective drugs, such as insulin and penicillin, are "magic bullets," which target diseases with a high degree of specificity and effectiveness. Unfortunately, very few magic bullets exist, because of many facts about the complex pathophysiology of diseases and the ways that exogenous drugs interact with our physiology (Stegenga 2015). Furthermore, most of the new drug applications submitted to the FDA are "me-too" drugs—drugs that are very similar to pre-existing drugs and that often have trifling effectiveness. A good example of "me-too" drugs are selective serotonin reuptake inhibitors: there are many members of this class of drugs, they bring their manufacturers great profit, and they are barely effective (Angell 2004; Kirsch et al. 2008).

Indeed, there is reason to think that the opposite of the concern expressed by Becker is true. Profit for pharmaceutical companies can be had by effective marketing rather than effective drugs—low regulatory standards can bring profit to companies whether or not their products are truly effective, precisely because low regulatory standards can be met by products with little effectiveness. If the FDA increased its epistemic standards, the profit incentive would remain, so in response pharmaceutical companies could be spurred to develop more effective drugs. In short, views like that expressed by Becker are unreasonably optimistic about the value of new pharmaceuticals, and demanding that research on new pharmaceuticals meet higher epistemic standards would not hinder an otherwise productive pipeline of effective drugs, and indeed might even enhance the development of more effective drugs.

A related counterargument to the thesis presented here is that drug development is already very costly, and increasing the epistemic standard for drug approval will further increase the cost of drug development. This cost would be passed on to patients, and since many drugs are already very expensive, the thesis presented here will make the expense of drugs even more burdensome. Some estimates hold that new drugs, on average, cost over $500 million to get FDA approval (cited in Resnik 2007). Others argue that this estimate is grossly inflated because the estimate includes corporate activity which is better thought of as marketing rather than research and development (Angell 2004). In any case, there is a cost associated with getting FDA approval for new drugs, and the counterargument to my thesis is that rendering the FDA's inductive risk calculus more balanced will add more cost. This counterargument is unconvincing for a number of reasons. Perhaps most important, it is not solely the cost of drugs which matters to patients or to payers (government healthcare systems or private insurers in the United States). Payers and consumers ultimately care about a more complicated property of drugs than simply cost, namely, the benefit accrued to the patient due to the effectiveness of a drug relative to the financial cost of the drug and the harms caused by the drug. In order to properly assess this more complex property, we must have more and better evidence regarding the effectiveness and harmfulness of drugs. Furthermore, many of the proposals suggested in the previous section for modulating the FDA's inductive risk calculus, such as the requirement of trial registration or the employment of appropriate measurement instruments, are relatively simple suggestions that would not add significant costs to drug development. Further, the concern about cost to consumers is misguided, since the bulk of the expense of new drugs is a result of the temporary monopoly granted to manufacturers of new drugs thanks to the patent system—new pharmaceuticals typically are very expensive because their manufacturers can charge whatever they want without competition from other manufacturers during the period in which the new pharmaceutical is protected by patent.

There is a growing movement to speed up the drug approval process, and an extreme example of this movement is a class of state-level laws that allow patients with life-threatening diseases access to experimental drugs that have not yet been approved by the FDA (Napier-Pearce 2015). The FDA already has a compassionate use clause, which allows for access to experimental drugs in particular circumstances. Similar bills have been passed by some states, which greatly reduces the amount of government oversight in granting such access to experimental pharmaceuticals. At first glance, such laws sound attractive—who could be opposed to such "compassionate use" clauses, which allow access to potentially life-saving drugs for patients with terminal illnesses? However, the matter is not so straightforward. First, novel effective medicines are extremely rare, certainly much rarer than most people suppose, and this dearth of effective medicines is not a result of regulation but rather is a result of the complex nature of diseases and the ways that pharmaceuticals act in our body. For terminal diseases, effective medicines are rarer still. Thus, it is typically not the case that strong pharmaceutical regulation keeps patients with terminal diseases from accessing life-saving drugs because the vast majority of the time such drugs simply do not, and cannot, exist. Second, such "compassionate use" movements should be assessed in the broader context surrounding the politics of federal regulation. The state-level bills permitting access to experimental drugs not yet approved by the FDA have been initiated by the Goldwater Institute, a conservative and libertarian organization (named after the former Senator Barry Goldwater) explicitly opposed to federal regulation. These bills are attempts to chip away at federal regulatory authority and are only secondarily concerned with patients' access to drugs (Napier-Pearce 2015). One might respond to this by holding that terminally ill patients have nothing to lose and thus should be free to try anything, but in fact terminally ill patients, like all people, have much to lose by consuming experimental interventions (foremost, the quality of their remaining life).

My argument supports the growing view in philosophy of science that non-epistemic values play a role in setting standards of evidence. This is especially salient in policy contexts such as drug regulation. The drug approval process illustrates the importance of exploring the full range of consequences when determining the appropriate standards of evidence (both good and bad consequences, following Elliott 2011), from a variety of perspectives (Wylie 1992). Non-epistemic values can and must determine standards of evidence in policy contexts, and there are, at least sometimes, good reasons (based on sociological, political, or scientific considerations) to employ particular value judgments when setting standards of evidence. A regulator's position on an inductive risk calculus is a proper subject of rational evaluation and can be more or less justified by ethical, political, and scientific considerations. In this chapter, I have argued that the inductive risk calculus for drug approval is skewed too far toward the extreme of avoiding unwarranted drug rejections. This inductive risk calculus should be retuned to be more balanced—this could be achieved by increasing the epistemic standards for assessing new drug applications.

Acknowledgments

I am grateful to Aaron Kenna, Ted Richards, and Kevin Elliott for detailed commentary.

References

Angell, Marcia. 2004. *The Truth about the Drug Companies: How They Deceive Us and What to Do about It*. New York: Random House.

Becker, Gary S. 2002. "Get the FDA Out of the Way, and Drug Prices Will Drop." *Bloomberg Business*, September 15.

Biddle, Justin B. 2013. "Institutionalizing Dissent: A Proposal for an Adversarial System of Pharmaceutical Research." *Kennedy Institute of Ethics Journal* 23(4): 325–53.

Carozza, Dick. 2005. "FDA Incapable of Protecting U.S., Scientist Alleges." *Fraud Magazine*, September/October.

CDER. 1998. "Guidance for Industry: Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products." Edited by Food and Drug Administration US Department of Health and Human Services, Center for Drug Evaluation and Research.

Decullier, Evelyne, An-Wen Chan, and François Chapuis. 2009. "Inadequate Dissemination of Phase I Trials: A Retrospective Cohort Study." *PLoS Medicine* 6(2): e1000034. doi: 10.1371/journal.pmed.1000034.

Douglas, Heather E. 2000. "Inductive Risk and Values in Science." *Philosophy of Science* 67(4): 559–79.

Douglas, Heather E. 2009. *Science, Policy and the Value-Free Ideal.* Pittsburgh, PA: University of Pittsburgh Press.

Dwan, Kerry, Douglas G. Altman, Jaun A. Arnaiz, Jill Bloom, An-Wen Chan, Eugenia Cronin, Evelyne Decullier, et al. 2008. "Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias." *PLoS ONE* 3(8): e3081. doi: 10.1371/journal.pone.0003081.

Elliott, Kevin C. 2011. "Direct and Indirect Roles for Values in Science." *Philosophy of Science* 78(2): 303–24.

Elliott, Kevin C., and Daniel J. McKaughan. 2009. "How Values in Scientific Discovery and Pursuit Alter Theory Appraisal." *Philosophy of Science* 76(5): 598–611.

Eyding, Dirk, Monika Lelgemann, Ulrich Grouven, Martin Härter, Mandy Kromp, Thomas Kaiser, Michaela F. Kerekes, Martin Gerken, and Beate Wieseler. 2010. "Reboxetine for Acute Treatment of Major Depression: Systematic Review and Meta-Analysis of Published and Unpublished Placebo and Selective Serotonin Reuptake Inhibitor Controlled Trials." *BMJ* 341. doi: 10.1136/bmj.c4737.

Friedman, Milton, and Rose Friedman. 1990. *Free to Choose: A Personal Statement*. New York: Mariner Books.

Goldacre, Ben. 2012. *Bad Pharma: How Drug Companies Mislead Doctors and Harm Patients.* New York: Farrar, Straus and Giroux.

Institute of Medicine. 2006. "The Future of Drug Safety: Promoting and Protecting the Health of the Public." Institute of Medicine. https://www.nap.edu/catalog/11750/the-future-of-drug-safety-promoting-and-protecting-the-health.

Kantrowitz, Arthur. 1978. "In Defense of the Science Court." *Hastings Center Report* 8(6): 4. doi: 10.2307/3561458.

Katz, Russell. 2004. "FDA: Evidentiary Standards for Drug Development and Approval." *NeuroRx* 1(3): 307–16.

Kirsch, Irving, B. J. Deacon, T. B. Huedo-Medina, A. Scoboria, T. J. Moore, and B. T. Johnson. 2008. "Initial Severity and Antidepressant Benefits: A Meta-Analysis of Data Submitted to the Food and Drug Administration." *PLoS Medicine* 5(2): e45. doi: 10.1371/journal.pmed.0050045.

Kitcher, Philip. 2011. *Science in a Democratic Society*. New York: Prometheus Books.

Krimsky, S. 2003. *Science in the Private Interest*. Lanham, MD: Rowman and Littlefield.

Leibovici, Leonard. 2001. "Effects of Remote, Retroactive Intercessory Prayer on Outcomes in Patients with Bloodstream Infection: Randomised Controlled Trial." *BMJ* 323(7327): 1450–1.

Lurie, Peter, and Sidney Wolfe. 1998. "FDA Medical Officers Report Lower Standards Permit Dangerous Drug Approvals." Public Citizen. http://www.citizen.org/Page.aspx?pid=2339.

McKeown, Thomas. 1976. *The Modern Rise of Population*. London: Edward Arnold.

Napier-Pearce, Jennifer. 2015. "Ethics of 'Right to Try' Bill for Experimental Drugs." *Salt Lake Tribune*.

Reiss, Julian. 2010. "In Favour of a Millian Proposal to Reform Biomedical Research." *Synthese* 177(3): 427–47. doi: 10.1007/s11229-010-9790-7.

Resnik, David. 2007. *The Price of Truth: How Money Affects the Norms of Science*. New York: Oxford University Press.

Rudner, Richard. 1953. "The Scientist qua Scientist Makes Value Judgements." *Philosophy of Science* 20: 1–6.

Stegenga, Jacob. 2015. "Measuring Effectiveness." *Studies in the History and Philosophy of Biological and Biomedical Sciences* 54: 62–71.

Stegenga, Jacob. 2016. "Hollow Hunt for Harms." *Perspectives on Science* 24: 481–504.

Stegenga, Jacob. Forthcoming. *Medical Nihilism*. Oxford: Oxford University Press.

Wilholt, Torsten. 2009. "Bias and Values in Scientific Research." *Studies in History and Philosophy of Science Part A* 40(1): 92–101. doi: http://dx.doi.org/10.1016/j.shpsa.2008.12.005.

Wylie, Alison. 1992. "The Interplay of Evidential Constraints and Political Interests: Recent Archaeological Research on Gender." *American Antiquity* 57(1): 15–35. doi: 10.2307/2694833.