The Routledge Companion to Philosophy of Medicine

Miriam Solomon, Jeremy R. Simon, Harold Kincaid

Measuring Harms

Publication details
Jacob Stegenga
Published online on: 25 Oct 2016

How to cite:- Jacob Stegenga. 25 Oct 2016, Measuring Harms from: The Routledge Companion to Philosophy of Medicine Routledge
Accessed on: 09 Aug 2023

PLEASE SCROLL DOWN FOR DOCUMENT

Full terms and conditions of use: https://www.routledgehandbooks.com/legal-notices/terms

This Document PDF may be used for research, teaching and private study purposes. Any substantial or systematic reproductions, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The publisher shall not be liable for an loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
MEASURING HARMS

Jacob Stegenga

Introduction

The benefits and harms of pharmaceuticals are principal subjects of investigation in clinical science. In this chapter I discuss how harms are measured and note the challenges that clinical research faces in detecting harms of pharmaceuticals. This chapter provides an introduction to the structure of clinical research with a focus on harm detection. As it is usually performed today, clinical research does not reliably measure the harms of pharmaceuticals. There are at least three categories of problems with clinical research that lead to the underestimation of the harm profile of pharmaceuticals: subtle features of research methodology, secrecy surrounding the evidence from clinical research, and inadequate regulation.

Clinical research is organized into four phases. Phase 1 trials evaluate an experimental drug in a small number of subjects to see if any very serious harms result from the drug. Phase 2 trials are usually randomized controlled trials in several hundred subjects and are preliminary evaluations of the benefits of an experimental drug. Phase 3 trials are larger randomized controlled trials, with several hundred to several thousand subjects, in which the potential benefits of the experimental drug are precisely measured. Approval of the experimental drug for general use occurs after Phase 3 trials. Once the drug is being used in general clinical settings, Phase 4 trials are sometimes done to gather more evidence about the benefits and harms of the drug.

Several important problems impede the detection of harms of experimental drugs in clinical research. Even before any evidence is gathered, the ways that harms are defined and planned to be measured in clinical research contributes to their underestimation. Although harms of experimental drugs are initially evaluated in Phase 1 trials, evidence from these trials is rarely published, although such trials provide the foundation for assessing the harm profile of pharmaceuticals. Phase 2 and 3 randomized controlled trials (RCTs) are usually designed to be as sensitive as possible to detecting potential beneficial effects of pharmaceuticals, but not sensitive enough to detecting rarer potential harmful effects of pharmaceuticals. This is especially troublesome given that RCTs are usually thought to produce the best evidence for causal hypotheses in medicine (although the epistemological status of RCTs is a matter of debate; see Chapters 18, 19, and 20). Phase 4 surveillance of harms contributes to underestimating the harm profile of pharmaceuticals, because harms are under-reported in general clinical settings. At every phase of medical research, the evidence regarding harms of pharmaceuticals is shrouded in secrecy (as is the evidence regarding benefits), which further contributes to the underestimation of the harm profiles of new drugs (and the overestimation of benefits). I use several examples to illustrate the problems with detecting harms of pharmaceuticals, and one
example in particular—rosiglitazone (trade name: Avandia) and the research about its harm profile—runs throughout the chapter.

The net effect of these conceptual, methodological, and social factors is that our available pharmaceuticals appear to be safer than they truly are. Were these factors mitigated in medical research, the harm profile of pharmaceuticals would be more faithfully represented, and pharmaceuticals would be deemed more harmful than they now are.

**Defining and Measuring Harms**

A harm of a pharmaceutical is, of course, an effect of the intervention, just as a benefit is an effect of the intervention. The interpretation of an effect as a harm (or conversely, as a benefit) is a normative judgment, and such is influenced by social values. Such judgments are not always straightforward. A compelling illustration is provided by the drug methylphenidate (Ritalin), which is frequently prescribed to treat attention deficit hyperactivity disorder (ADHD). The alleged benefits of methylphenidate depend on a particular social nexus and are conceptually intertwined with its potential harms. Empirical tests of methylphenidate suggest that it mitigates bodily motions of children diagnosed with ADHD and mitigates frequency of social interactions, which might be seen as a benefit by a school teacher. But this effect could be seen as harmful by someone who thinks that children moving around, playing, and socializing are generally positive behaviors, with a large range of normal or healthy levels. Thus, the same effect of a pharmaceutical may be considered a benefit or a harm depending on one’s broader normative commitments and sociocultural position. There are, though, many effects of pharmaceuticals that can be considered harms with little ambiguity in typical cases, from insubstantial effects such as a minor headache to more severe effects such as death.

Harmful effects of drugs are often thought of as discrete outcomes, referred to as adverse events, or if they are extremely harmful, serious adverse events. Harms, however, can be changes of continuous parameters (and not just changes of discrete parameters); many harms should not be thought of as events, since discrete events constitute only a subset of the potential harms of a pharmaceutical. As I discuss below, much data about harms of drugs come from passive surveillance after a drug has been approved for general use. Passive surveillance can only identify a token event as a harm if a patient or a physician observes and interprets the event as a harm and reports it as such. Small effects and common effects are often not reported. If a drug causes a patient to gain several pounds, this effect might go unnoticed by the patient and the physician, and even if it were noticed, there is no way that the patient or the physician could reliably assess the drug as a cause of the weight gain. In other words, such an effect might not be attributed as an effect of the drug, and if it were, such an attribution would not be reliable.

Terminological choices also contribute to obscuring the harm profiles of pharmaceuticals. Concern about harms of pharmaceuticals are often referred to with terms such as “drug safety” (indeed, the discipline sometimes referred to as “pharmacovigilance” is also referred to as drug safety). A report of a new kind of harm of a pharmaceutical is a “signal” of a “safety finding” (!), which is documented via “safety reporting.” For example, the FDA, when talking about serious harms such as death and strokes caused by peroxisome proliferator activated receptor (PPAR) modulators (described below), referred to these events as “clinical safety signals,” and some drugs in this class were removed from the market because of “clinical safety.” We should beware of use of the term “safety” to refer to harm.

Beyond mere terminological choices, the way harms are defined and measured in trials contributes to their underestimation. For example, before it was established that some
antidepressants cause suicidal ideation and behavior, meta-analyses of clinical trial data sug-
ggested that these drugs do not in fact cause such harms. The data in these trials came from
patient outcomes measured with the Hamilton Rating Scale for Depression (HAMD). The
HAMD is a poor instrument for measuring benefits of antidepressants, because interventions
with effects that are irrelevant to the core elements of depression (such as mitigation of fidget-
ing or a slight change in sleeping patterns) can contribute to a large change in HAMD score
(Stegenga, 2015). But the HAMD is an even worse instrument for measuring harms of antide-
pressants, including suicidality. There is one question on the HAMD regarding suicidality, as
follows: “Suicide. 0 = Absent. 1 = Feels life is not worth living. 2 = Wishes he were dead or any
thoughts of possible death to self. 3 = Suicidal ideas or gesture. 4 = Attempts at suicide (any
serious attempt rates 4).” The numbers refer to the points contributed to the overall HAMD
score (the higher the score, the more severely depressed a patient is said to be). The problem is
that an antidepressant could cause a patient who has occasional but passing suicidal thoughts
to develop severe and frequent suicidal ideation and self-mutilation, yet the patient's HAMD
score would not change, since both before and after the antidepressant the patient would
receive a score of 3 on suicidality. The HAMD is insensitive to such harms of antidepressants.
Since the HAMD is the measuring instrument employed in research on antidepressants, such
research systematically underestimates the harms of antidepressants. This example illustrates
the general point that the way harms are defined and measured in research can contribute to
their underestimation.

Another example of how defining and measuring harms contributes to their underestima-
tion is from the drug rosiglitazone (Avandia), once the world’s best-selling drug for type 2
diabetes. By 2007, evidence was mounting that rosiglitazone causes cardiovascular disease and
death. GlaxoSmithKline, the manufacturer of rosiglitazone, funded a large trial (the RECORD
trial) in an attempt to show that rosiglitazone was safe. The primary outcome measured in the
RECORD trial was a composite outcome that included all hospitalizations and deaths from
any cardiovascular causes. Although this outcome measure appears reasonable given the goal
of the study, it in fact is quite problematic. The problem is that it included hospitalizations that
were very likely not related to the randomized interventions. Since we can presume that hospi-
talizations and deaths that were not caused by the interventions occurred at roughly the same
rate between the different groups, the overall larger number of this outcome in both groups
minimized the relative difference in outcomes observed between the groups (in other words,
the important outcome of interest—cardiovascular disease and death caused by rosiglitazone—
was watered down, making it less likely that the study would detect a statistically significant
difference between the groups). Moreover, the outcome hospitalization depends, obviously, on
a patient being hospitalized, but this is a socio-economic decision as much as a health-related
outcome; the trial included patients in dozens of countries, including many in eastern Europe,
with diverse hospitalization practices. This diversity of practice could have introduced vari-
ability into the data, which would make it more likely that a statistically significant difference
between experimental groups would not be detected. In short, the way that the potential harm
was defined and measured in this trial artificially lowered the chance of detecting the harm of
the drug in question.

The broader point illustrated by the examples above is that harms of pharmaceuticals will
only be found if the harms are properly defined and measured. Defining and measuring a harm
in the wrong ways—such as by employing a measuring instrument or an outcome that is insen-
tive to the harm in question—amounts to not looking for the harm.

The definition and measurement of harms influences (and constrains) the evidence that
is gathered at each phase of clinical research. I now turn to the particular phases of clinical
research.
MEASURING HARMs

Phase 1: First in Human

A first-in-human study is an experiment in which a pharmaceutical is first tested in humans. Experimental pharmaceuticals are often initially evaluated with cell and animal experiments, and if such experiments provide evidence that suggests that the pharmaceutical is relatively safe and potentially effective for humans, then a first-in-human study is performed, usually in healthy volunteers. These are also referred to as Phase 1 trials.

Such trials are obviously risky for the research subjects. Despite the risk, Phase 1 trials are important because they provide the foundation for assessing harms of pharmaceuticals. Given that first-in-human studies are the first time a potential new pharmaceutical is tested in humans, they provide crucial evidence regarding harms. Such evidence is relevant, obviously, to the harm profile of the particular molecule under investigation, but it is also relevant to the harm profile of the class of molecules to which the particular molecule belongs and is more broadly relevant to the harm profile of drugs, generally. Evidence from first-in-human studies is, therefore, extremely important.

Unfortunately, such evidence is rarely shared publicly. The vast majority of Phase 1 trials are not published (Decullier, Chan, & Chapuis, 2009). Publication bias in clinical research refers to the phenomenon in which studies that suggest that a drug is beneficial and/or safe are more likely to be published than those studies that do not. Those molecules that appear to be relatively safe in a Phase 1 trial often go on to be tested in Phase 2 and Phase 3 trials, and thus the broader scientific community can infer that such molecules have passed a Phase 1 trial and so are at least somewhat safe. Those molecules that appear to be relatively harmful in Phase 1 trials hardly ever go on to be tested in later-phase trials, and results from such Phase 1 trials are rarely published. This failure to publish Phase 1 trials is wasteful, because future scientists who become interested in the same compound may needlessly repeat Phase 1 trials they were unaware of (trial registries have been proposed to address publication bias of Phase 3 trials, but thus far have not been very helpful, and I am not aware of trial registries for Phase 1 trials).

Beyond being wasteful, this needless repetition also has the potential for causing needless harm to subjects in these subsequent trials. There is, though, a further consequence of not publishing Phase 1 trials that is even more widespread and sinister.

The problem is that when assessing the harm profile of an experimental pharmaceutical, if one is unaware of past evidence regarding harms, then one’s prior assessment that the molecule is harmful will be lower than it should be (that is, lower than it would be if one was aware of such evidence). Since molecules that appear safe in Phase 1 trials go on to be evaluated in larger and more public Phase 2 and 3 trials, and molecules that appear harmful do not, and since most Phase 1 trials are not published, it follows that, of all drugs that are approved for clinical use, the proportion that appears harmful is lower, perhaps much lower, than is truly the case. In other words, publication bias of Phase 1 trials is systematically skewed: publicly available evidence from Phase 1 trials tends to suggest that experimental drugs are safe, but unavailable evidence from Phase 1 trials tends to suggest that experimental drugs are harmful. Since we have access to the former but not the latter, our general assessment of the harms of pharmaceuticals is grossly underestimated.

Returning to our example, rosiglitazone is a modulator of proteins called peroxisome proliferator-activated receptors (PPARs), which regulate the expression of genes. In recent years, more than 50 PPAR modulators have failed clinical tests, and many of these failures have been due to harms caused by the PPAR modulators (Nissen, 2010). Indeed, evidence of such harms was available even prior to first-in-human studies: for example, PPAR modulators were found to cause numerous types of tumors and cardiac toxicity in rodent studies. Unfortunately, “few publications have detailed the precise toxicity encountered” (Nissen, 2010).
In addition to past evidence regarding harms of experimental drugs from Phase 1 trials, another factor that ought to influence one's assessment of the probability that an experimental drug is harmful is background knowledge of the way the experimental drug intervenes in normal and pathological physiological mechanisms. PPAR modulators are again a good example: any given PPAR modulator can influence the expression of many dozens of genes, and thus “the effects of these agents are unpredictable and can result in unusual toxicities” (Nissen, 2010). Unfortunately, given the emphasis on RCTs in clinical research, this kind of knowledge is often downplayed. Indeed, mechanistic reasoning is typically denigrated by contemporary evidence-based medicine. Taking such knowledge into account, even if such knowledge were incomplete, would render our estimations of the harm profiles of pharmaceuticals more accurate.

Given this argument, we should expect examples of drugs that appear to be relatively safe based on evidence from Phase 1 trials, but then come to be viewed as relatively harmful based on evidence from clinical trials and post-market surveillance (Phases 2–4). This is precisely what we observe. Just among the class of PPAR modulators, troglitazone has been withdrawn in some jurisdictions because it appears to cause liver damage; tesaglitazar has been withdrawn in some jurisdictions because it appears to cause elevated serum creatinine; pioglitazone has been withdrawn in some jurisdictions because it appears to cause bladder cancer; and mura- glitazar has been withdrawn in some jurisdictions because it appears to cause heart attacks, strokes, and death.

In short, the lack of availability of evidence from Phase 1 trials contributes to the systematic underestimation of harms of pharmaceuticals.

**Phases 2 and 3: Randomized Controlled Trials**

The next steps in testing an experimental pharmaceutical are Phase 2 and Phase 3 RCTs. Although these are larger than Phase 1 trials, as typically employed in clinical research, RCTs are not good methods for detecting harms of pharmaceuticals. In principle, RCTs could be more reliably employed to hunt for harms, though such trials would have to be larger and longer than most trials performed today and incorporate other more fine-grained methodological changes. In practice, RCTs are designed to be sensitive to the detection of benefits of pharmaceuticals instead of being sensitive to the detection of harms.

Trial designers try to maximize the ability of a trial to detect the potential benefits of an experimental pharmaceutical; they employ a number of design strategies to do so. One is to maximize the observed effect size in an RCT by including only subjects who are most likely to show the most benefit of the intervention in question. For example, many trials testing antidepressants include only the most severely depressed patients, and antidepressants have only been shown to work for such patients. Another trial design strategy is to minimize the potential variability of data generated by trials—for example, RCTs usually include a relatively homogeneous group of subjects. The greater the similarity among subjects in a trial with respect to parameters that can influence outcomes of a trial (such as age, sex, or the presence of other diseases), the less variable the data. Moreover, RCTs usually exclude subjects who have diseases other than the one being intervened on in the trial, who are on other pharmaceuticals, and who are elderly. The most egregious examples of inclusion and exclusion criteria are called “enrichment strategies.” Enrichment strategies involve excluding subjects—after subject recruitment but prior to the collection of hard data—who respond especially well to placebo or who respond especially poorly to the experimental pharmaceutical. Such strategies are widely employed in clinical research.
The net effect of these strategies is that subjects of trials are different in many important respects from the patients who might use new pharmaceuticals once they are approved for use in a clinical setting. Some of these features that often differ between research subjects and clinical patients are known to influence the harm profiles of pharmaceuticals. For example, older people, pregnant women, and patients on other drugs are more likely to be harmed by pharmaceuticals, but they are also precisely the kinds of people who are excluded from RCTs. For example, the most common harm of statins is myopathy, which ranges from simply myalgia (muscle pain) and muscle weakness, to rhabdomyolysis, a severe condition in which muscle tissue dies and releases proteins into the blood, which can cause kidney failure and death. This risk is higher among women, elderly people, and people with other conditions like infections, seizures, and kidney disease—precisely the kinds of people that are excluded from clinical trials.

In short, the exclusion of certain kinds of patients and inclusion of other kinds of patients entails that the sensitivity of RCTs to detect harms is typically much less than the sensitivity of RCTs to detect benefits. The subjects who are included in a trial are less likely to be harmed by an experimental pharmaceutical than an average person is, and more likely to benefit. In another striking example, Worrall (2010) notes that in the large ASSENT-2 trial, one exclusion criterion was “any other disorder that the investigator judged would place the patient at increased risk.” Of course, there is a basis for such an exclusion criterion—namely, the protection of patients who are more liable to be harmed by experimental pharmaceuticals. However, this exclusion criterion directly reduced the ability of the trial to detect the harms of the experimental pharmaceutical that would result when it would be employed in a real-world clinical setting (since it is precisely in the clinical setting in which patients have other disorders that put them at increased risk of harm).

To return to rosiglitazone, a meta-analysis showed that rosiglitazone causes an increased risk of heart attack and death (Nissen & Wolski, 2007), although the RCTs in the meta-analysis were too small individually to have adequate power to detect this rare but severe harm. GlaxoSmithKline funded the RECORD trial in an attempt to refute this finding. RECORD employed seven inclusion criteria and 16 exclusion criteria, and 99% of the subjects were Caucasian. One effect of these criteria was that subjects in the trial were, on average, healthier than the broader population; for example, subjects in both the experimental group and control group of the trial had a heart attack rate of close to half that in the equivalent demographic group in the broader population. Thus, the results from this trial were less applicable to the target population than it would have been had the trial not employed such restrictive exclusion and inclusion criteria.

In addition to restrictive exclusion and inclusion criteria, a further reason why RCTs are typically not very good at finding harms is that there is often insufficient measurement of a wide range of physiological parameters that could indicate that the pharmaceutical causes harm. In principle, a pharmaceutical could negatively impact any of the countless physiological parameters of a subject, but these harms will not be detected if they are not measured. Moreover, even if such parameters are measured, they are often not reported. A survey of 142 randomly selected publications of RCTs of psychiatric interventions found that only a fraction bothered to address harms, and on average, reports of RCTs used a small fraction of a page in the results section to discuss harms (Papanikolaou, Churchill, Wahlbeck, & Ioannidis, 2004). In other words, harms that are not looked for will not be found.

Two other limitations of RCTs contribute to the underestimation of harms of pharmaceuticals: their size and their duration. RCTs normally enroll enough subjects to detect the potential benefit of the pharmaceutical for the disease in question. Any further subjects add
expense. However, this number of subjects is often not enough to detect harms that are rare. The duration of a trial is normally also just long enough to detect the potential benefit of the pharmaceutical for the disease in question. Some RCTs only evaluate pharmaceuticals for a period of weeks, as did many of the trials of rosiglitazone. However, some harms of drugs manifest only after years of taking the drug. Methylphenidate (Ritalin), for example, has been shown to cause stunted growth in children, but this is only found three years after the initiation of treatment with the drug (Swanson et al., 2007). For these reasons, larger and longer Phase 4 observational studies are usually relied on to detect harms (but as I note below, Phase 4 studies have their own practical and epistemic shortcomings).

Since RCTs underestimate harms, we should expect to observe examples of drugs that appear relatively safe after RCTs but come later to appear more harmful once used in a clinical setting. This phenomenon is widespread. In the worst-known cases, pharmaceuticals are pulled from the market by manufacturers or regulators, such as valdecoxib (Bextra), fenfluramine (Fondamine), gatifloxacins (Gatiflo), and rofecoxib (Vioxx) over just the last few years. In other cases, the harm profile in the clinical setting appears greater than RCTs suggested, but the drugs have been left on the market for whatever reason (often, regulators consider the benefit-harm profile of the drug to remain favorable despite the increased estimation of its harm profile). A few examples include celecoxib (Celebrex), alendronic acid (Fosamax), risperidone (Risperdal), and olanzapine (Zyprexa).

**Phase 4: Post-Approval**

The majority of data on harms of pharmaceuticals come from observational studies and passive surveillance conducted after a pharmaceutical has been approved for clinical use. These studies are sometimes called Phase 4 studies. The fact that the majority of data regarding harms of pharmaceuticals comes from post-market studies has an important practical consequence, and the fact that such studies are usually observational designs has an important epistemic consequence.

The bar for approval of a new pharmaceutical is low. The FDA, for example, requires only that two Phase 3 RCTs show some benefit, regardless of how many RCTs were performed, and although such RCTs are usually not very sensitive to the harm profile of the pharmaceutical. After the new pharmaceutical has been approved for clinical use, the potential harms of the pharmaceutical are assessed by passive surveillance systems and observational studies. Although the harm profile of new pharmaceuticals are not yet known at this point, they are nonetheless prescribed to and consumed by patients, often numbering in the millions. Only when the new pharmaceuticals are used in a clinical setting (rather than an experimental setting) are most data on harms gathered. These data come from patients who are prescribed the drug by their physician and thus unwittingly become subjects in a study of the harm profile of the drug.

There is strong reason to think that, like the earlier phases of drug evaluation, post-market passive surveillance also significantly underestimates harms of pharmaceuticals. One empirical evaluation puts the underestimation rate at 94% (Hazell & Shakir, 2006). Part of the reason for this is that there is widespread under-reporting of possible harmful effects (as noted above, patients have to identify harmful effects, and patients and physicians have to infer that the harmful effect is indeed a result of a drug, and the physician must then take the time to report the harmful effect).

A few large-scale randomized trials have included a thorough hunt for harms, and Papanikolaou, Christidi, and Ioannidis (2006) compared estimates of harms from these trials to equivalent non-randomized trials. They found that non-randomized studies, on average,
have conservative estimates of harms of pharmaceuticals relative to randomized trials of comparable sizes on the same intervention. One reason cited for such a finding is that patients who reliably take their prescribed medications on schedule tend to be healthier than non-compliant patients, and thus there is a confounding factor when comparing the outcomes of those patients who consume more of a particular pharmaceutical with those patients who consume less of the pharmaceutical. That is, those who consume more of a medication (those who are faithful to their medication schedule) also tend to be healthier than those who take less, so observational studies tend to overestimate the benefits of pharmaceuticals and underestimate their harms.

The kind of evidence most often available in post-approval research can influence our estimations of the harm profiles of drugs. Most evidence regarding harms of pharmaceuticals comes from non-randomized studies, and the dominant view of the evidence-based medicine (EBM) movement is that non-randomized studies are unreliable. Therefore, the EBM movement denigrates the majority of evidence regarding harms of pharmaceuticals. This is important because EBM is now a very influential movement—for instance, EBM proponents’ views about RCTs has influenced regulators such as the FDA. According to this line of thinking, only RCTs can provide compelling evidence regarding harms of pharmaceuticals. Since the majority of data regarding harms comes from non-randomized studies, and since the data regarding harms that does come from RCTs is fundamentally limited for the reasons noted above, regulators are liable to make unreliable judgments regarding the harms of pharmaceuticals.

Vandenbroucke (2008) and Osimani (2014) have argued against the view that RCTs are better than non-randomized (e.g., observational) studies at detecting harms. Because harms of drugs are unintended and often unknown effects, physicians cannot bias treatment allocation with respect to such effects. Thus, so-called selection bias is less of an epistemological worry for unintended harmful effects as it is for intended beneficial effects, and so one of the central advantages of RCTs over observational studies is mitigated in the context of the hunt for harms.

**Secrecy**

Evidence regarding harms of pharmaceuticals is shrouded in secrecy. Companies that pay for RCTs claim that they own the data from the trials, and clinical researchers who participate in RCTs are bound by gag clauses in their contracts that constrain their ability to share data, even if they suspect that the pharmaceutical under investigation causes harm. One form of secrecy is the ubiquity of publication bias (discussed above about Phase 1 trials). Just like first-in-human studies, Phase 3 trials suffer from substantial publication bias, resulting in the benefits of novel pharmaceuticals being exaggerated and the harms of novel pharmaceuticals being underestimated. Another form of secrecy involves the withholding of data from independent researchers by both corporate manufacturers and government agencies. As with first-in-human studies, if regulators do not have access to all of the available data, then they cannot make a reliable inference regarding the harms of drugs.

As one example, reboxetine is an antidepressant (SSRI) sold in Europe during the past decade. Recently, researchers with access to both published and unpublished data performed a meta-analysis of the drug’s effects (Eyding et al., 2010). Of the 13 trials that had been performed on reboxetine, data from 74% of patients remained unpublished. Seven of the trials compared the drug to placebo: one had positive results and only this one was published; the other six trials (with almost 10 times as many patients as the positive trial) showed no benefit for reboxetine, 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 worse than its competitors on the primary outcome and had worse side effects. (For more details on this episode, see Goldacre, 2012.) Thus, just as with Phase 1 trials, Phase 3 trials suffer from publication bias.

Once one starts looking for examples of the secrecy surrounding evidence of harms of pharmaceuticals, they become easy to find. Olanzapine (Zyprexa) is now known to cause extreme weight gain and concomitant diabetes, but the manufacturer, Eli Lilly, “engaged in a decade-long effort to play down the health risks of Zyprexa according to hundreds of internal Lilly documents and e-mail messages among top company managers” (Berenson, 2006). Oseltamivir (Tamiflu) provides yet another example: evidence on the harms of oseltamivir largely remain unpublished, despite the massive stockpiling of the drug by Western countries in recent years (Doshi, 2009).

Rosiglitazone, again, provides a striking example. After several trials suggested that rosiglitazone may cause cardiovascular harms, Nissen and colleagues requested this trial data from GlaxoSmithKline, which refused to share it. However, due to a lawsuit regarding secrecy about another of its drugs (Paxil), the company had agreed to develop a registry of data from its clinical trials, and Nissen was able to access this data. These investigators identified 42 RCTs of rosiglitazone, of which only seven had been published. The resulting meta-analysis, which included the unpublished studies, showed that rosiglitazone increases cardiovascular harms by 43%. Within 24 hours of the meta-analysis being submitted to the New England Journal of Medicine, one of the peer reviewers faxed a copy of the manuscript to GlaxoSmithKline. Internal company emails discussed the similarity of Nissen’s findings to GlaxoSmithKline’s own analysis, which they had performed years earlier but had not published. Moreover, the FDA had performed their own analysis, which reached similar conclusions, but also did not publicize its findings. The director of research at the company wrote: “FDA, Nissen, and GSK all come to comparable conclusions regarding increased risk for ischemic events, ranging from 30% to 43%!“ (Harris, 2010). In other words, the FDA and GlaxoSmithKline already knew of the cardiovascular harm caused by rosiglitazone, but neither the regulator nor the company had publicized this finding.

When the secrecy of evidence about harms of pharmaceuticals is threatened by vigilant researchers, manufacturers occasionally respond belligerently. Rosiglitazone, again, provides a good illustration. John Buse, one of the world’s foremost diabetes researchers, gave two talks in 1999 arguing that rosiglitazone may have cardiovascular risks. GlaxoSmithKline then executed an orchestrated campaign to silence Dr. Buse. This plan appears to have been initiated by the company’s head of research, and even the chief executive officer was aware of it. The company referred to Dr. Buse as the “Avandia Renegade,” and in contact with Buse and Buse’s department chair there were “implied threats of lawsuits.” The company’s head of research wrote in an internal email:

I plan to speak to Fred Sparling, his former chairman as soon as possible. I think there are two courses of action. One is to sue him for knowingly defaming our product even after we have set him straight as to the facts—the other is to launch a well planned offensive on behalf of Avandia. . . .

Buse responded to the company with a letter that ended by capitulating and asking them to “please call off the dogs.” (This episode was detailed in a U.S. Senate Finance Committee Report, Anon, 2007.) Later, Buse expressed embarrassment that he caved to the pressure of GlaxoSmithKline. Only with the publication of Nissen’s meta-analysis in 2007 was the potential of rosiglitazone to cause ischemic events revealed. By that time, the FDA estimated, rosiglitazone had caused about 83,000 heart attacks since coming on the market in 1999.
Conclusion

Various solutions have been proposed to address some of the problems of detecting harms of pharmaceuticals. There are some obvious candidates, including increasing the quality of evidence in the hunt for harms, improving the accessibility of such evidence, and improving regulation of clinical research. The harm profile of a pharmaceutical is, obviously, necessary in order to evaluate the benefit-harm balance of the pharmaceutical. Because harms of pharmaceuticals are systematically underdetected at all stages of clinical research, policy makers and physicians generally cannot adequately assess the benefit-harm balance of pharmaceuticals.

References


Further Reading


Osimani, B. (2014). Hunting side effects and explaining them: Should we reverse evidence hierarchies upside down? *Topoi, 33*(2), 295–312. (An argument that the evidence-based medicine evidence hierarchies should be turned “upside down” for assessing the harms of pharmaceuticals)

Worrall, J. (2002). What evidence in evidence-based medicine? *Philosophy of Science, 69*, S316-S330. (By now a classic article that invigorated the study of the epistemology of medical research)