# Hollow Hunt for Harms

**Jacob Stegenga**
*University of Victoria*
*University of Johannesburg*

*Harms of medical interventions are systematically underestimated in clinical research. Numerous factors—conceptual, methodological, and social—contribute to this underestimation. I articulate the depth of such underestimation by describing these factors at the various stages of clinical research. Before any evidence is gathered, the ways harms are operationalized in clinical research contributes to their underestimation. Medical interventions are first tested in phase 1 "first in human" trials, but evidence from these trials is rarely published, despite the fact that such trials provide the foundation for assessing the harm profile of medical interventions. If a medical intervention is deemed safe in a phase 1 trial, it is tested in larger phase 2 and 3 clinical trials. One way to think about the problem of underestimating harms is in terms of the statistical power of a clinical trial—the ability of a trial to detect a difference of a certain effect size between the experimental group and the control group. Power is normally thought to be pertinent to detecting benefits of medical interventions. It is important, though, to distinguish between the ability of a trial to detect benefits and the ability of a trial to detect harms. I refer to the former as power$_B$ and the latter as power$_H$. I identify several factors that maximize power$_B$ by sacrificing power$_H$ in phase 3 clinical trials. If a medical intervention is approved for general use, it is evaluated by phase 4 post-market surveillance. Phase 4 surveillance of harms further contributes to underestimating the harm profile of medical interventions. At every stage of clinical research the hunt for harms is shrouded in secrecy, which further contributes to the underestimation of the harm profiles of medical interventions.*

## 1. Introduction

Harmful effects of medical interventions are systematically underestimated by clinical research. Such underestimation is a product of conceptual,

I am grateful to Alex Broadbent, Jonathan Fuller, Heather Douglas and two anonymous reviewers for commentary on earlier drafts. For discussion of the ideas presented here I am grateful to Beatrice Golomb, Eran Tal, Martin Carrier, Nancy Cartwright, and audiences at University of Toronto, University of Utah, Aarhus University, Bielefeld University, University of Redlands, and the Brocher Foundation.

Perspectives on Science 2016, vol. 24, no. 5
 doi:10.1162/POSC_a_00220

methodological, and social factors present throughout the various stages of clinical research.

The difficulties with detecting harms of medical interventions begin with how harms are conceptualized and operationalized in clinical research. Harms are often thought of as discrete outcomes, referred to as adverse events, which are often measured by the same methodological apparatus employed to measure benefits. Harms, however, can be more subtle (and more widespread) than discrete outcomes, and indeed, generally ought not to be thought of as events at all. The ways that harms are operationalized in clinical research contributes to underestimating their severity and frequency (§2).

After preliminary research on cells, tissues, and animals, phase 1 "first in human" studies are performed to evaluate the harm profile of novel medical interventions. Unfortunately the vast majority of such studies remain unpublished, which generally and systematically skews the overall assessment of harm profiles of medical interventions (§3). If a medical intervention is deemed safe in a phase 1 trial, it is tested in larger phase 2 and 3 clinical trials. Randomized controlled trials (RCTs) constitute one of the most significant hurdles in the hunt for harms. Most RCTs are designed to be sensitive to the detection of potential benefits of medical interventions, and this sensitivity trades-off against the sensitivity to detect potential harms of medical interventions. This is especially troublesome given that RCTs are usually thought to produce the best evidence for causal hypotheses in medicine. In §4 I highlight several ways that RCTs are designed such that the harms of interventions under investigation are underestimated. Once a medical intervention has been approved for use in the clinical setting, harms are hunted with the use of passive surveillance and sometimes with phase 4 trials. This has both practical and epistemic shortcomings (§5).

The hunt for harms is embedded in a social nexus that exacerbates the underestimation of harms. Most evidence regarding the harms of medical interventions is generated by studies which are funded and controlled by the manufacturers of the interventions under investigation, and whose interests are best-served by underestimating the harm profile of such interventions. This leads to widespread limitation of the evidence regarding harms that is made available to independent scientists and policy-makers, and this, in turn, contributes to the underestimation of the harm profiles of medical interventions. Regulators lack the authority to properly estimate harm profiles of medical interventions, and frequently contribute to shrouding the relevant evidence regarding harms in secrecy (§6).

The net effect of these conceptual, methodological, and social factors is that our available medical interventions appear to be safer than they truly are. Were these factors mitigated in medical research, the harm profiles of

medical interventions would be more faithfully represented, and medical interventions would be deemed more harmful than they now are.

## 2. Operationalizing Harm

A harm of a medical intervention is, of course, an effect of the intervention, just as a benefit of an intervention is an effect. The interpretation of an effect *as* a harm (or conversely, as a benefit) is a normative judgment, and as such is influenced by social values. Such judgments are not always straightforward. A compelling illustration is provided by the drug methylphenidate (Ritalin), often 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 children's bodily motions and frequency of social interactions, which might be seen as a benefit by an overworked teacher. But this effect could be judged as harmful by someone who thinks that children moving around, playing, and socializing are generally positive behaviors. As one critic puts it, stimulants like methylphenidate "work for the teacher" (Whitaker 2010). Unfortunately, evidence suggests that methylphenidate does not work for the student. Self-reports of well-being and assessments of academic performance are not improved by methylphenidate, and in the longer term, methylphenidate causes worse outcomes (see §4). Thus the same effect of a medical intervention may be considered a benefit or a harm depending on one's broader normative commitments and sociocultural position. There are, though, many effects of medical interventions 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.

Harms are often thought of as discrete outcomes, referred to as "adverse events," or if they are extremely harmful, "serious adverse events." A harm, however, can be a small change of a continuous parameter rather than a change of a discrete parameter; many harms should not be thought of as events, since discrete events constitute only a fraction of the potential harms of a medical intervention. As I discuss in §5, the vast majority of harm data comes from passive surveillance and observational studies, in which a particular token event can only be detected as a harm in the first place if a patient or a physician observes and interprets an effect of a medical intervention as a possible harm, and reports it as such. Small effects and common effects are often not reported. If a drug causes a patient to gain two pounds, this effect could easily 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 a minor effect probably would not be

attributed as an effect of the drug, and if it were, such an attribution would not be reliable.[1] Nevertheless, if the drug were consumed by millions of people in an already overweight society, such an effect could have profoundly negative consequences.

Terminological choices contribute to obscuring the harm profiles of medical interventions. Concerns about harms of medical interventions are often referred to with terms such as "drug safety" (indeed, the object of the study of drug harms is sometimes called "drug safety"). A report of a new kind of harm of a medical intervention is a "signal" of a "safety finding" (!), which is documented via "safety reporting." For example, when talking about serious harms such as death and strokes caused by peroxisome proliferator activated receptor (PPAR) modulators (described below), the FDA referred to these events as "clinical safety signals," and some drugs in this class were removed from the market because of "clinical safety." The use of the term *safety* to refer to *harm* is perhaps the most egregious Orwellian locution in medicine. Moreover, other benign-sounding phrases employed in discourse about medical interventions, such as "side effects" or "adverse events"—which can refer to collapsing lungs, self-mutilation, exploding tendons, and death—contribute to the opacity of harms of medical interventions.

The way harms are operationalized contributes to their underestimation. For example, before it was well established that antidepressants can cause suicidal ideation, some analyses of clinical trial data suggested that these drugs do not in fact cause this terrible harm. The data from these trials were of patient outcomes measured with the Hamilton Rating Scale for Depression (HAMD). Elsewhere I argue that the HAMD is a poor instrument for measuring the efficacy of antidepressants, because interventions with effects that are irrelevant to core elements of depression (such as mitigation of fidgeting or a slight change in sleeping patterns) can contribute to a large change in HAMD score (Stegenga, forthcoming-a, forthcoming-c). But the HAMD is an even worse instrument for measuring certain harms of antidepressants, 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, according to the usual interpretation of the

1. This also holds, in principle, for small effects that are beneficial. However, as I argue in §4, clinical trials are typically designed to be more sensitive to detecting benefits than they are to detecting harms.

scale). 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 without actually attempting suicide, and 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 primary measuring instrument employed in RCTs of antidepressants, such RCTs systematically underestimate harms of antidepressants. This example illustrates the more general point that the way harms are operationalized in clinical research can contribute to the underestimation of the harm profile of medical interventions.

Another example of how the operationalization of harms contributes to their underestimation is from the drug rosiglitazone (Avandia), once the world's best-selling drug for type-2 diabetes (I will return to this example throughout this article). By 2007, evidence was mounting that rosiglitazone causes cardiovascular disease and death (several journalists have written about this case; see, for example, Goldacre 2012). GlaxoSmithKline, the manufacturer of rosiglitazone, funded a large trial (the RECORD trial), in an attempt to disprove this. The primary outcome measured in the RECORD trial was a composite outcome that included all hospitalizations and deaths from any cardiovascular causes. This gerrymandered outcome included cardiovascular hospitalizations that were very likely not related to the randomized interventions (rosiglitzaone or control), and thus, because we can presume that hospitalizations and deaths that were not caused by either intervention occurred at roughly the same rate between the trial groups, the overall larger number of this gerrymandered outcome in both groups minimized the relative difference in outcomes observed between the groups (the important outcome of interest—cardiovascular disease and death caused by rosiglitzaone—was, in both groups, "watered down" by including the much more frequent outcome of hospitalization, making it less likely to detect a statistically significant difference between the groups). Moreover, the outcome "hospitalization" depends on, obviously, a patient being hospitalized, but this is a socio-economic decision as much as a health-related outcome, and the trial included patients in dozens of countries with diverse hospitalization practices. This diversity of practice could have introduced variability in 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 operationalized in the RECORD 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 medical interventions will only be found if properly looked for. Operationalizing a harm in certain ways—such as by employing a measuring instrument

or an outcome which is insensitive to the harm in question—amounts to not looking for the harm.

### 3. First-in-Human, Never Seen Again

A first-in-human study is an experiment in which a medical intervention is administered in humans for the first time. Generally, medical interventions first are evaluated with *in vitro* and animal experiments, and if such experiments provide evidence to think that the medical intervention might be safe and potentially effective for human use, a first-in-human study is performed. Such studies are also referred to as phase 1 trials.

Such studies carry significant risks for the subjects of the trials. A recent first-in-human study of a molecule called CD28-SuperMAB (also referred to as TGN1412) tested a dose of the experimental drug that was 500 times lower than the dose found to be safe in animals.[2] The six men who were given the drug quickly developed intense headaches, back pain, intestinal pain, diarrhea, fevers, low blood pressure, or lung pain, and after 48 hours each had multiple organ failures.

Despite the risk of first-in-human studies, they are important because they provide the foundation for assessing harms of medical interventions. Given that first-in-human studies are the first time an experimental medical intervention 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. I will use the following notation in the arguments to come: molecule $x$ is a member of the class of molecules of type $T$, and this class is itself a member of the class of all drugs $D$. Evidence from a first-in-human study on $x$ is relevant, obviously, to the harm profile of $x$, but is also relevant to the harm profile of $T$ (albeit more indirectly), and is also relevant to the harm profile of $D$ (more indirectly still). Evidence from first-in-human studies is, therefore, hugely important.

Unfortunately, such evidence is rarely shared publicly. The vast majority of first-in-human studies are not published (for empirical evidence on this, see Decullier, Chan, and Chapuis 2009). It is difficult to know exactly what proportion of first-in-human studies are published because there is no registry of what molecules have been tested by first-in-human studies.

2. Maël Lemoine has written an insightful analysis of this case (presently unpublished). The dosage given to the human subjects relative to the safe dosage in animals was in fact an estimate based on extrapolation from an animal model of CD28-SuperMAB, since this class of drug is species-specific.

Empirical studies of the publication bias of first-in-human trials have relied on records of institutional review boards (committees of universities and hospitals for reviewing proposed experiments involving humans). Cases like CD28-SuperMAB, in which the public learns about the harm profile of a molecule because of a tragedy in the respective first-in-human study, are atypical. For experimental interventions with less dangerous harm profiles we know very little about such harm profiles because the evidence regarding the harm profile of the vast majority of molecules is rarely published or publicized.

Those molecules that appear to be relatively safe in a first-in-human study 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 first-in-human study and so are at least somewhat safe. Those molecules that appear to be relatively harmful in first-in-human studies rarely go on to be tested in further trials, and such first-in-human studies are rarely published. This publication bias of first-in-human studies is wasteful. Future scientists who are unaware of the harm profile of $x$ or other molecules of class $T$, for which prior first-in-human studies have been performed, and who want to know the harm profile of $x$ or another member of $T$, are liable to perform wasteful subsequent first-in-human studies. This also has the potential for causing needless harm to subjects in these subsequent first-in-human studies. There is, though, a consequence of publication bias of first-in-human studies that is much more widespread and sinister.

When assessing the harm profile of a molecule ($x$), if one is unaware of past evidence regarding harms (of $x$ and more generally $T$), then one's prior probability that the molecule is harmful will be lower than it should be (that is, lower than it otherwise would be if one was aware of such evidence). Since molecules that appear safe in first-in-human studies tend to be evaluated in larger and more public phase 2 and 3 studies, and molecules that appear harmful do not, and since most first-in-human studies are not published, it follows that, of all drugs that are tested for clinical use, the proportion that *appears* harmful is lower, perhaps much lower, than is truly the case.

This is crucially important, so I reiterate the argument in more formal terms. Our assessment of the harm profile of $x$ can be represented as a conditional probability, $P(H|E)$, where H is the hypothesis that $x$ is harmful, and E is relevant new evidence regarding the harm profile of $x$ (E could be data from a first-in-human study, or from a phase 3 trial, or whatever). The conditional probability can be rearranged according to Bayes' Theorem as:

$$P(H|E) = P(E|H)P(H) / P(E)$$

Thus, our assessment of the harm profile of a molecule is directly proportional to P(H), the prior probability that the molecule is harmful. How ought one determine P(H)? This is a notoriously difficult question for scientific methodology. But in this context there is at least an obvious constraint on an answer. The prior probability that $x$ is harmful depends on past evidence regarding $x$ and other molecules like it, including molecules of type $T$ and more broadly all drugs $D$. We have access to only a small subset of the relevant past evidence regarding harms of $x$ and $T$ and $D$. Given the rampant publication bias noted above, the evidence that we do not have access to tends to confirm H more often than does the evidence that we do have access to. It follows that our assessment of P(H) would be higher if we had access to all relevant evidence. This, concomitantly, would have a direct positive impact on P(H|E) (as seen in Bayes' formula above). Thus, *for all drugs*, our estimate of the probability that any particular drug is harmful is artificially lower than it otherwise would be if we had all the relevant evidence from first-in-human studies. The extent of this problem is difficult to estimate, but given the empirical estimates of the frequency of publication bias of first-in-human trials noted above, it appears devastating.

What is the appropriate reference class for assessing the harm profile of $x$? Like the question about assessing P(H), this is a notoriously difficult question for scientific methodology. Again, though, in this context there is a straightforward constraint on an answer. Since $x$ has some close similarities with other molecules of type $T$, and broad similarities with other members of $D$, when assessing the harm profile of $x$ at the very least one should take into account the harm profiles of more members of $T$ and $D$ than is currently possible due to publication bias. And since, as above, the publication bias regarding the harm profiles of $T$ and $D$ is systematically skewed toward underestimating harms, it follows that if one were able to assess the harm profile of $x$ with a more appropriate reference class, the harm profile of $x$ would be more accurately assessed and $x$ would appear more harmful than it otherwise does.

Returning to my running 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, according to a leading type-2 diabetes researcher, "few publications have detailed the precise toxicity encountered" (Nissen, 2010), and "few data on toxicity are available

in the public domain because of the common industry practice of not publishing safety findings for failed products" (Nissen and Wolski, 2007).

Another factor that ought to influence one's assessment of the prior probability that $x$ is harmful is background knowledge of the way that $x$ 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 randomized controlled trials (RCTs) in clinical research, this kind of knowledge is often downplayed. Indeed, mechanistic reasoning is typically denigrated by contemporary evidence-based medicine (see Illari 2011).

Given the argument presented in this section, one would expect to see examples of drugs which 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). And of course, this is precisely what we observe. Just among the class of PPAR modulators there are many such examples: 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 muraglitazar has been withdrawn in some jurisdictions because it appears to cause heart attacks, strokes, and death.[3]

Publication bias of other study types further contributes to the systematic underestimation of the harms of medical interventions, as I discuss in §6. In short, the lack of availability of evidence from first-in-human studies contributes to the systematic underestimation of harms of medical interventions.

### 4. Clinical Trials and the Abuse of Power

Clinical trials, as typically employed in biomedical research, are not good methods for hunting harms of medical interventions. As Rawlins puts it: "in the assessment of harms RCTs are weak at providing evidence … they are an unreliable approach to the definitive identification of harms" (2008, p. 15). In this section I argue that although Rawlins is correct about the unreliability of RCTs to detect harms of medical interventions, this is a contingent shortcoming of RCTs caused by particular fine-grained decisions

3. By "withdrawn" here I mean that the particular drug has been removed from a national jurisdiction based on the noted harm. Some of these drugs are still available in some jurisdictions.

regarding methodological designs of most clinical trials. In principle RCTs could be 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 medical interventions at the expense of being sensitive to the detection of harms.

To make this argument I employ the concept of statistical power. The statistical power of a clinical trial is characterized in several ways: the probability that a statistical analysis of data from a trial rejects a false null hypothesis; the probability of avoiding a "type II" error (falsely concluding that there is no difference between the experimental group and the control group); and the probability of detecting a difference between the experimental group and the control group if there is truly a difference to be detected. Broadly construed, power refers to the sensitivity of a trial to detect an effect of the intervention under investigation, when there is such an effect to be detected. The power of a trial depends on three parameters: the effect size of the intervention under investigation, the number of subjects in the trial, and the variability of the data. It is usually difficult to achieve satisfactory power in trials of novel medical interventions, for a variety of reasons: so many novel medical interventions have relatively small effects, increasing the number of subjects in a trial is expensive, and research subjects can respond in very different ways to experimental interventions.

Trial designers try to maximize power in a number of ways. One is to maximize the observed effect size in a trial by including only subjects who are most likely to show the most benefit of the intervention in question. For example, some trials testing antidepressants include the most severely depressed patients—and generally antidepressants have been shown to work only in such patients. The parameter that influences power, which is often easiest for trial designers to control, is the variability of the data: to minimize data variability, trial designers include a relatively homogeneous group of subjects in the trial. The greater the similarity among subjects in a trial with respect to parameters that are known to often influence outcomes of a trial (such as age, sex, or the presence of other diseases), the less variable will be the data from the trial. Finally, despite the expense, trials will often include many thousands of subjects. In short, trial designers employ several strategies to maximize power. (There is, obviously, great financial incentive to avoid the error of falsely concluding that a potential new medical intervention is ineffective.)

To maximize the observed effect size and minimize the variability of data, trial designers employ various criteria constraining what subjects are included or excluded from the trial. For example, it is typical to exclude elderly subjects, subjects on other drugs, or subjects with other diseases.

The most egregious of these trial design features are called "enrichment strategies": after the enrollment of subjects, but prior to the start of data collection, subjects are tested for how they respond to placebo or the experimental intervention, and those subjects that do well on placebo or (and sometimes and) those subjects that do poorly on the experimental intervention are excluded from the trial.[4]

One effect of these strategies is that subjects of trials are different in many relevant respects from the patients who use new medical interventions once they are approved for use in a clinical setting. Some of these differences are themselves known to influence the harm profiles of medical interventions. Older people, pregnant women, and patients on other drugs (for example) are more likely to be harmed by a novel medical intervention, but they are also precisely the kinds of people who are excluded from trials. For example, the most common harm of statins is myopathy, which ranges from simple muscle pain and weakness, to rhabdomyolysis, which is a severe condition in which muscle tissue dies and releases proteins (myoglobin) into the blood, which can cause kidney failure and death (other harms of statins include stroke, congenital defects, diabetes, cancer, neuromuscular symptoms, nerve damage, abnormal liver function, joint problems, and tendon damage). 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. One aim of excluding such patients is to maximize power—that is, to maximize the ability to detect potential benefits of a drug—but the exclusion of such patients also minimizes the ability to detect potential harms of a drug.

Here is a striking example. Worrall notes that in the large ASSENT-2 trial, an exclusion criterion was "any other disorder that the investigator judged would place the patient at increased risk" (2010, p. 297). Of course, there is a basis for such an exclusion criterion, namely, the protection of patients who are more likely to be harmed by experimental interventions. However, this exclusion criterion directly mitigated the ability of

4. One type of enrichment strategy is called a "run-in" period, which involves the exclusion of placebo-responders before the trial begins. Here is an example. Of 15 trials recently analyzed by the FDA regarding antidepressant use in children, only three showed positive results. Two of these three studies were of fluoxetine (Prozac), and thus fluoxetine was approved for use in children diagnosed with depression. However, the trials put all children on a placebo for one week, and any children who significantly improved during this week were excluded from the trial. This trial employed this run-in period in order to maximize the difference between the expected beneficial effects among children in the fluoxetine group compared to children in the placebo group, but in so doing, rendered the subjects in the trial notably different from real-world patients.

the trial to detect the harms of the intervention that would result when the intervention were 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. The enrichment strategy which involves excluding subjects who fare poorly on the test drug is another example of a trial design feature which directly mitigates the ability of a trial to detect harms of medical interventions, because those subjects most likely to experience harms caused by the interventions are excluded from the trial.

Usually, trial power refers to the ability of a trial to detect a benefit. However, since a harm of a medical intervention is simply another effect of the medical intervention, just like a benefit, the power of a trial can also refer to the ability of a trial to detect harms. I will call the sensitivity of a trial to detect benefit "$\text{power}_B$" and the sensitivity of a trial to detect harm "$\text{power}_H$." $\text{Power}_B$ and $\text{power}_H$ trade-off against each other. There are numerous ways in which trial designers attempt to maximize $\text{power}_B$ at the expense of $\text{power}_H$. The exclusion of certain kinds of patients and inclusion of other kinds of patients, mentioned above, is one such strategy. The net result is that the power of trials (and more broadly the sensitivity of trials) to detect harms is typically much lower than the power of trials to detect benefits. For an empirical demonstration of this, (Tsang, Colley, and Lynd 2009) performed their own calculations of the statistical power to detect serious harms of medical interventions. The original trial publications that this group analyzed did not report the power to detect harms, though the trials did report that no statistically significant serious harms were found. When Tsang and colleagues calculated $\text{power}_H$ for the trials, they found values ranging from 0.07 to 0.37. This means that the probability is very high that these trials would falsely report that there are no harms of the medical interventions in question even if there were in fact harms.

To return to my running example, a meta-analysis was performed which showed that rosiglitazone causes an increased risk of heart attack and death from cardiovascular disease (Nissem and Wolski 2007). The individual trials that this group amalgamated were too small to have an adequate power to detect this rare but severe harm. GlaxoSmithKline funded the RECORD trial in an attempt to show that rosiglitazone does not increase the risk of heart attacks. This trial employed seven inclusion criteria and 16 exclusion criteria, and 99% of the subjects were Caucasian. One result of these criteria was that subjects in the trial were, on average, healthier than the broader population; for example, subjects in the trial (that is, in both the control group and the rosiglitzaone group) had a heart attack rate about 40% less than the heart attack rate in the equivalent demographic group (middle-aged people with type-2 diabetes) in the broader population (and it was heart attack that was the very harm in question).

If subjects in the experimental group of a trial withdraw from the trial due to harms of the medical intervention under investigation, then the presentation of the trial data could give a misleading impression that the medical intervention is safer than it actually is, because data about harms in those subjects who withdrew is not collected. Unfortunately there is scant evidence about the frequency of subject withdrawal. Insufficient reporting of subject withdrawals is ubiquitous. For example, in a review of 133 publications of RCTs published in 2006 in six leading general medical journals, Pitrou et al. (2009) found that no information on severe adverse events was given in 27% of the articles, and no information on subject withdrawal due to adverse events was given in 47% of the articles.

There are two other limitations of clinical trials that are widely recognized as factors that contribute to the underestimation of harms of medical interventions: their size and their duration. Clinical trials normally enroll enough subjects to detect the potential benefit of the medical intervention for the disease in question. Any more subjects add expense. However, this number of subjects is often not enough to detect harms that are severe but rare. Trial size is optimized to achieve satisfactory $\text{power}_B$, without concern for $\text{power}_H$. The duration of a trial is normally also just long enough to detect the potential benefit of the medical intervention for the disease in question. Some studies of antidepressants, for example, only evaluate the drugs for a period of weeks, as did many of the trials of rosiglitazone. A longer trial adds expense. 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 (by as much as 2 cm in height and 2.7 kg in weight), but this is only found three years after the initiation of treatment with the drug (Swanson et al. 2007). For these two reasons—the small size and short duration of trials—larger and longer passive observational studies are usually relied on to detect harms (but as I note below, passive observational studies have their own practical and epistemic shortcomings).

Since clinical trials underestimate harms, we should expect to observe examples of drugs which appeared to be relatively safe after clinical trials but came later to appear to be more harmful once used in a clinical setting. This phenomenon is widespread. The worst cases are those in which manufacturers or regulators pull medical interventions from the market. Here are a few examples from the last several years: valdecoxib (Bextra), fenfluramine (Pondimin), gatifloxacin (Gatiflo), and rofecoxib (Vioxx). Other cases are those in which the harm profile in the clinical setting appears worse than RCTs suggested, but have been left on the market for whatever reason (often, regulators consider the benefit-harm profile of the drug to remain favorable regardless of the increasing estimation of its harm profile). A few examples include: celecoxib (Celebrex), alendronic acid

(Fosamax), risperidone (Risperdal), olanzapine (Zyprexa), and in the United States, the running example of this paper, rosiglitazone (Avandia). Some of these drugs have been the subject of massive lawsuits because the manufacturers deceptively downplayed the harm profiles of the drugs. Another important property of trials that contributes to the underestimation of harms of medical interventions is not intrinsic to the methodology of the trials themselves, but is rather about how the evidence from the trials is shared publicly and used by regulators. I explore this in §6.

In short, clinical trials are not usually well designed to hunt for harms.[5] A survey of 142 randomly selected reports of clinical trials of psychiatric interventions found that only a fraction bothered to address harms, and on average, reports of trials used 1/10 of a page in the results section to discuss harms (Papanikolaou et al. 2004). In this section I have argued that two important methodological properties of clinical trials—$\text{power}_B$ and $\text{power}_H$—trade-off against each other, usually in favor of $\text{power}_B$ at the expense of $\text{power}_H$, and this tradeoff is constituted by a plurality of fine-grained methodological choices made by trial designers.

How this trade-off between $\text{power}_B$ and $\text{power}_H$ is balanced is obviously influenced by non-epistemic values, such as the financial value associated with avoiding the error of concluding that an experimental intervention is more harmful than it truly is or the social value associated with avoiding the error of concluding that an experimental intervention is less harmful than it truly is. Douglas (2000) gives a prominent articulation of the thesis that non-epistemic values can influence scientific inference.

## 5. Jump Now, Look Later (But Don't Look Hard)

The vast majority of data regarding harms of medical interventions comes from observational studies and passive surveillance conducted after a given medical intervention has been approved for clinical use. These studies are sometimes called "phase 4" post-market studies. The fact that the majority of data regarding harms of medical interventions 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 that a new medical intervention has to get over in order to be approved for consumption and marketing is low. The FDA, for example, requires only two RCTs that show that a new medical intervention has some beneficial effect, regardless of how many RCTs were performed on the new medical intervention, and despite the fact that the $\text{power}_H$ of such RCTs is usually extremely low, and thus, at the point at which a new medical intervention is approved for general use, there is scant evidence available on

5. Meta-analyses of RCTs are no better, for reasons given in Stegenga, 2011.

the harm profile of that intervention. After the new medical intervention has been approved for clinical use, the potential harms of the medical intervention are assessed by passive surveillance systems and observational studies. With little evidence available regarding the harm profile of new medical interventions, such interventions are prescribed and consumed by typical patients, often numbering in the millions. It is only at this point, when the new medical interventions are used in clinical settings, rather than an experimental setting, that most data on harms is gathered. This data comes from patients who are prescribed the latest drug by their physician and who inadvertently become subjects in a study regarding the harm profile of the drug. Without knowing it, such patients are unwitting guinea pigs in the hunt for harms.

There is strong reason to think that post-market passive surveillance severely underestimates harms of medical intervention. One empirical evaluation of this puts the underestimation rate at 94% (this was based on a wide-ranging empirical survey by Hazell and Shakir 2006).

Unfortunately, because observational studies and passive surveillance do not involve a randomized design, they are typically denigrated relative to randomized controlled trials. For instance, a methodological textbook claims that "if a study wasn't randomised, we suggest that you stop reading it and go on to the next article in your search" (Straus et al. 2005, p. 58). Since most evidence regarding harms of medical interventions comes from non-randomized studies (especially rare severe harms), the dominant view of the evidence-based medicine (EBM) movement thereby denigrates the majority of evidence regarding harms of medical interventions.[6]

This view has influenced regulators. For instance, this passages comes from testimony from an employee of twenty years at the FDA, now the associate director of the FDA's Office of Drug Safety (there's that word again!), during a congressional hearing regarding the drug rofecoxib (Vioxx):

> The corporate culture within CDER [Center for Drug Evaluation and Research, part of the FDA] is also a barrier to effectively protecting the American people. The culture is dominated by a world-view that believes only randomized clinical trials provide useful and actionable information and that postmarketing safety is an afterthought.[7]

6. There is a growing literature criticizing the standard EBM view regarding the relative merits of observational and randomized studies; among many others, see Worrall 2002, Cartwright 2007, Bluhm 2009, Borgerson 2009, and Bluhm 2010.

7. The full testimony, which is a scathing account of the ineffective regulation provided by the FDA, is available at: http://www.finance.senate.gov/imo/media/doc/111804dgtest.pdf (accessed June 30, 2015).

According to the line of thinking criticized in this testimony, only RCTs can provide compelling evidence regarding harms of medical interventions, and 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 I described above, U.S. regulators, by their own lights, have a paucity of reliable evidence regarding harms of medical interventions.

Vandenbroucke (2008) has articulated an argument that, in the context of hunting for harms, RCTs are better than observational studies. 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 (see also Osimani 2014). The upshot is that observational studies do not typically *overestimate* harm profiles of drugs. Indeed, there is some empirical evidence suggesting that observational studies *underestimate* harm profiles.

There have been a few large-scale RCTs that included a thorough hunt for harms, and Papanikolaou et al. (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 medical interventions relative to RCTs of comparable sizes on the same intervention. One reason cited for such a finding is that those patients who take their prescribed medications on a schedule which is faithful to their physicians' orders 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 medical intervention compared with those patients who consume less of a particular medical intervention (namely, those who consume more of a medication, given that they are faithful to their medication schedule, also tend to be healthier than those who take less). Thus, observational studies that compare those patients who consume more medical interventions than other patients tend to overestimate the benefits of medical interventions and underestimate their harms. For the above reasons, even if regulators do not typically have access to evidence regarding harms of medical interventions from large-scale RCTs, they could rely on evidence from comparable non-randomized studies and be confident that, on average at least, they are not overestimating the harm profile of medical interventions.

The evidence regarding harms that regulators do have access to, however, is apparently good enough to keep secret, as I explore in the next section.

### 6. Secrecy of Data

A vast amount of evidence regarding harms of medical interventions is shrouded in secrecy. Companies that pay for clinical trials claim that they own the data from the trials, and clinical researchers that participate in industry-sponsored trials are often bound by gagging clauses in their contracts that constrain their ability to share data for any reason, even if they suspect that a particular medical intervention under investigation causes harm.

Consider reboxetine, for example. Reboxetine is an antidepressant (SSRI) sold in Europe during the past decade. 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 (for more details of this case, see Goldacre 2012). Seven of the trials compared the drug against placebo: one had positive results and only this one was published; the other six trials (with almost ten times as many patients as the positive trial) 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 worse than its competitors on the primary outcome, and had worse side effects. Just like phase 1 first-in-human studies, phase 3 RCTs suffer from rampant publication bias, which results in the benefits of novel medical interventions being exaggerated and the harms being underestimated.

The tribulations of rofecoxib (Vioxx) provide a striking example of such secrecy, which was later publicly exposed. The manufacturer of rofecoxib, Merck, carried out the VIGOR trial to test the drug's safety and efficacy. It is now widely thought that in the landmark publication of the VIGOR trial (Bombardier et al. 2000), the authors withheld data on cardiovascular harms associated with rofecoxib. This was the view of the editors of the journal which published the article (*The New England Journal of Medicine*), after they learned of Merck memos that showed that at least two of the article's authors were aware of the data on cardiovascular harms (Curfman, Morrissey, and Drazen 2005). The methodological issue was portrayed by Merck and the article's authors as more subtle: the analysis of cardiovascular harms followed a predefined plan, according to which the study stopped collecting cardiovascular harm data on a particular date (Feb 10, 2000), and so they claimed that it would have been post hoc and thus inappropriate to include the reports of cardiovascular harms associated with rofecoxib that were gathered in the two weeks after this cutoff date (Bombardier et al.

2006).[8] This is controversial: many philosophers of science hold that the timing of when one gains a particular piece of evidence is irrelevant to how confirmatory that evidence is. Regardless, at the very least the cut-off date for gathering data on cardiovascular harms, and the particular evidence regarding cardiovascular harms that was in fact gathered after the cut-off date, could have been made public. Instead, such data were kept secret for too long.

Examples of the secrecy surrounding evidence of harms of medical interventions are easy to find. Here are three others. 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). Paroxetine (Paxil) provides another good example of egregious secrecy. GlaxoSmithKline (GSK), the manufacturer of paroxetine, hid evidence about the harmful effects of the drug for years. These harms include withdrawal symptoms and an increase in suicidality in children and teenagers. This secrecy led to a massive lawsuit.[9] Oseltamivir (Tamiflu) provides yet another example: evidence of the harms of oseltamivir largely remain unpublished, despite the massive stockpiling of the drug by Western countries in recent years (Doshi 2009).

The running example of this article, rosiglitazone, again provides a striking illustration of the secrecy surrounding evidence of harms of medical interventions. In this case, regulators themselves contributed to such secrecy. After several trials suggested that rosiglitazone may cause cardiovascular harms, Steve Nissen and his colleagues requested patient-level data from GlaxoSmithKline, which refused to share the data. But due to the lawsuit mentioned above regarding paroxetine, the company had agreed to develop a registry of data from their clinical trials. Nissen identified 42 RCTs of rosiglitazone, of which only seven had been published. The resulting meta-analysis showed that rosiglitazone increases cardiovascular events by 43%. Within 24 hours of submitting the meta-analysis

8. The cut-off date for reporting data on gastrointestinal events—the parameter which rofecoxib was thought to be superior to its competitors—was March 9, 2000, and thus the trial was more likely to gather data that suggested rofecoxib was superior to its competitor than it was likely to gather data that suggested rofecoxib was more harmful than its competitor. Thus the VIGOR trial provides another example of the thesis presented in §4. See Biddle 2007 for a detailed account of what he calls the "Vioxx debacle."

9. The systematic bias and fraud surrounding paroxetine is astonishing, and included millions of dollars in undisclosed payments from GSK to psychiatric researchers, deliberate withholding of evidence showing that paroxetine was ineffective in children, and rampant publication bias. For a dramatic book-length exposition of this case, see Bass 2008.

to the *New England Journal of Medicine*, one of the peer reviewers had faxed a copy of the manuscript to GlaxoSmithKline. Internal emails in the company discuss the similarity of Nissen's findings to their own analysis which they had performed years earlier but had not published. Moreover, the FDA had performed its own analysis which reached similar conclusions, but also did not publicize the 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%!" In short, the FDA and GlaxoSmithKline already had known of the cardiovascular harm caused by rosiglitazone, but neither the regulator nor the company had publicized this finding.

Indeed, regulators are not only often powerless against such secrecy, they are often complicit in it. Here is another example in which regulators were complicit in such secrecy. Cochrane researchers doing a systematic review of the diet drugs orlistat and rimonabant tried to get unpublished data from the European Medicines Agency (EMA) in June 2007. In August 2007 the EMA rejected the request for data, by invoking protection of commercial interests and intellectual property. The Cochrane researchers appealed to the European Union Ombudsperson, who found EMA to be guilty of maladministration, and found the EMA arguments for secrecy (patient confidentiality, commercial protection) to be unwarranted. Nevertheless, the EMA continued to withhold the evidence, and dismissed the argument that patients were at risk of harm. EMA claimed that the trial design principles in the relevant reports were commercially protected (!). Other researchers were also trying to get EMA documents on rimonabant. Finally the Cochrane researchers were sent 60 pages. However, these 60 pages were almost entirely redacted by the EMA.[10] (In 2009 rimonabant was taken off the market because it caused an increased risk of psychiatric problems and suicide.)

Sometimes the regulators are not complicit in such secrecy, they are simply inept. Oseltamivir again provides a striking example. When Cochrane researchers set out to update their systematic review of oseltamivir in 2009, they decided to include an assessment of the harm profile of the drug. They found that the FDA post-market Adverse Event Report System had fewer entries in total than Roche's own post-market surveillance system had for just neuropsychiatric harms. The Roche system listed 2466 neuropsychiatric events between 1999 and 2007, of which 562 were classified as "serious," while the FDA system only noted 1805 events of any kind (see Doshi 2009) for further discussion).

10. This case is discussed in Goldacre (2012). For the full document, see http://www.prescrire.org/editoriaux/edi33693.pdf (accessed June 30, 2015).

In his discussion of some of the above episodes, the physician and journalist Ben Goldacre asked regulators at the MHRA (British equivalent of FDA) and the EMA why they thought such secrecy was warranted. The response from staff at both agencies claimed that people outside these agencies are liable to misinterpret evidence regarding harms of medical interventions (they both mentioned the measles, mumps, and rubella vaccine scandal as an example of how such evidence can be misinterpreted).

When the secrecy of evidence on harms of medical interventions is threatened by vigilant researchers, manufacturers can respond belligerently. Rosiglitazone, again, provides a good illustration. Dr. John Buse, one of world's foremost diabetes researchers, gave two talks in 1999 arguing that rosiglitazone may have cardiovascular risks. GlaxoSmithKline executed an orchestrated campaign to silence 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 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…[11]

Buse responded to the company with a letter that ended by asking them to "please call off the dogs." Later Buse expressed embarrassment that he caved to the pressure of GlaxoSmithKline. By 2007, the year that Nissen's meta-analysis was published, the FDA estimated that rosiglitazone had caused about 83,000 heart attacks since coming on the market in 1999; the drug has been the subject of thousands of lawsuits. Dr. David Graham, a senior employee at the FDA whose testimony regarding rofecoxib I cited in §5, has called for rosiglitazone to be removed from the market. Despite this, the drug remains available in the United States.

What is wrong with secrecy of trial data? Most obviously, physicians, policy-makers, and patients cannot make informed treatment decisions if they do not have access to existing evidence on the harms of medical interventions.

11. This passage is cited in the US Senate Committee on Finance that released a report regarding the intimidation of Dr. Buse. Note that it was the Finance Committee, rather than a senate committee associated with the FDA, who initiated a congressional investigation. The full report, titled "The Intimidation of Dr. John Buse and the Diabetes Drug Avandia," was published in November 2007 and is available online.

Evidence from RCTs is, arguably, a public good that should be available for all to see, as clean water should be available for all to drink. Indeed Lemmens and Telfer (2012) argue that access to evidence from RCTs is a fundamental component of the right to health. Secrecy of evidence from clinical research impedes informed decisions (in the context of the present concern, by minimizing the apparent harms of medical interventions), and thereby frustrates what is arguably a fundamental human right.

## 7. Conclusion

The harm profile of a medical intervention is, obviously, necessary in order to evaluate the benefit-harm balance of the medical intervention. Because harms of medical interventions are systematically underestimated at all stages of clinical research, policy-makers and physicians generally cannot adequately assess the benefit-harm balance of medical interventions. Many seem to think that regulators can assess the benefit-harm balance. For example, Resnik expresses a widely held misperception about the ability of regulators to properly consider the harm profile of medical interventions:

> In gathering data, the company must adhere to FDA standards for clinical investigation, which address issues relating to subject safety and rights, research design, and data integrity. The FDA examines the data provided by the company from its clinical trials and balances the benefits and risks of the new product. The FDA will allow the company to market the product if it determines that the benefits of the product outweigh the risks. (Resnik 2007, p. 80)

The arguments in this article show just how misguided this view is.

Many medical interventions that are present-day blockbusters—consumed by many millions of patients—are not intended to treat diseases but are rather employed as preventive therapies. Statins to lower cholesterol and radiotherapy to avoid recurrence of local breast cancer are prominent examples. Such medical interventions typically have very low absolute effect sizes on patient-level outcomes (such as mortality), and thus have a high so-called "number needed to treat"—the number of patients that must use such a medical intervention in order to achieve a single positive patient-level outcome is very high. Thus, the harms of such medical interventions will be harms to patients the vast majority of whom do not receive a benefit of the medical intervention.

Various solutions have been proposed to address some of the problems of detecting harms of medical interventions. There are some obvious candidates, including increasing the quality of evidence in the hunt for harms, in order to increase $\text{power}_{\text{H}}$, and improving the accessibility of such evidence when it is available. One reason for the underestimation of harms

involves a fundamental trade-off between $power_B$ and $power_H$, and in this trade-off too frequently that which is sacrificed is the ability to detect harms of medical interventions. When evidence on harms exists, from the various stages of clinical research, it is often publicly unavailable. The harms of medical interventions remain systematically underestimated by clinical research.

**References**

Bass, A. 2008. *Side Effects: A Prosecutor, a Whistleblower, and a Bestselling Antidepressant on Trial*. Chapel Hill, NC: Algonquin Books.

Berenson, A. 2006. "Eli Lilly Said to Play Down Risk of Top Pill." *The New York Times*, December 17. http://www.nytimes.com/2006/12/17/business/17drug.html?_r=0

Biddle, J. 2007. "Lessons from the Vioxx Debacle: What the Privatization of Science Can Teach us about Social Epistemology." *Social Epistemology*, 21 (1): 21–39.

Bluhm, R. 2009. "Some Observations on Observational Research." *Perspectives in Biology and Medicine* 52 (2): 252–263. doi: 10.1353/pbm.0.0076.

Bluhm, R. 2010. "The Epistemology and Ethics of Chronic Disease Research: Further Lessons from ECMO." *Theoretical Medicine and Bioethics* 31 (2): 107–122. doi: 10.1007/s11017-010-9139-8

Bombardier, C., L. Laine, R. Burgos-Vargas, B. Davis, R. Day, M. B. Ferraz, C. Hawkey, M. C. Hochberg, T. Kvien, T. Schnitzer, and A. Weaver. 2006. "Response to Expression of Concern Regarding VIGOR Study." *New England Journal of Medicine* 354 (11): 1196–1199. doi: 10.1056/NEJMc066096.

Bombardier, C., L. Laine, A. Reicin, D. Shapiro, R. Burgos-Vargas, B. Davis, R. Day, M. Bosi Ferraz, C. Hawkey, M. C. Hochberg, T. Kvien, and T. J. Schnitzer. 2000. "Comparison of Upper Gastrointestinal Toxicity of Rofecoxib and Naproxen in Patients with Rheumatoid Arthritis." *New England Journal of Medicine* 343 (21): 1520–1528. doi: 10.1056/NEJM200011233432103.

Borgerson, K. 2009. "Valuing evidence: Bias and the Evidence Hierarchy of Evidence-based Medicine." *Perspectives on Biology and Medicine* 52 (2): 218–233. doi: 10.1353/pbm.0.0086.

Cartwright, N. 2007. "Are RCTs the Gold Standard?" *BioSocieties* 2 (1): 11–20. doi: 10.1017/s1745855207005029.

Curfman, G. D., S. Morrissey, and J. M. Drazen. 2005. Expression of Concern: Bombardier et al., "Comparison of Upper Gastrointestinal Toxicity of Rofecoxib and Naproxen in Patients with Rheumatoid Arthritis," *New England Journal of Medicine*, 2000, 343: 1520–1528. *New England Journal of Medicine* 353 (26): 2813–2814. doi:10.1056/NEJMe058314.

Decullier, E., A.-W. Chan, and F. Chapuis. 2009. "Inadequate Dissemination of Phase I Trials: A Retrospective Cohort Study." *PLoS Med*, 6 (2): e1000034. doi: 10.1371/journal.pmed.1000034.

Doshi, P. 2009. "Neuraminidase Inhibitors—The Story Behind the Cochrane Review." *British Medical Journal* 339: b5164. doi: 10.1136/bmj.b5164.

Douglas, H. 2000. "Inductive Risk and Values in Science." *Philosophy of science* 67 (4): 559–579.

Eyding, D., M. Lelgemann, U. Grouven, M. Härter, M. Kromp, T. Kaiser, M. F. Kerekes, M. Gerken, and B. 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." *British Medical Journal* 341. doi: 10.1136/bmj.c4737

Goldacre, B. 2012. *Bad Pharma: How Drug Companies Mislead Doctors and Harm Patients*. New York: HarperCollins.

Hazell, L., and S. A. Shakir. 2006. "Under-Reporting of Adverse Drug Reactions: A Systematic Review." *Drug Safety* 29 (5): 385–396.

Illari, P. M. 2011. "Mechanistic Evidence: Disambiguating the Russo–Williamson Thesis." *International Studies in the Philosophy of Science* 25 (2): 139–157. doi: 10.1080/02698595.2011.574856

Lemmens, T., and C. Telfer. 2012. "Access to Information and the Right to Health: The Human Rights Case for Clinical Trials Transparency." *American Journal of Law and Medicine* 38: 63–112.

Nissen, S. E. 2010. "The Rise and Fall of Rosiglitazone." *European Heart Journal* 31 (7): 773–776. doi: 10.1093/eurheartj/ehq016.

Nissen, S. E., and K. Wolski. 2007. "Effect of Rosiglitazone on the Risk of Myocardial Infarction and Death from Cardiovascular Causes." *New England Journal of Medicine* 356 (24): 2457–2471. doi: doi:10.1056/NEJMoa072761.

Osimani, B. 2014. "Hunting Side Effects and Explaining Them: Should we Reverse Evidence Hierarchies Upside Down?" *Topoi* 33 (2): 295–312.

Papanikolaou, P. N., G. D. Christidi, and J. P. Ioannidis. 2006. "Comparison of Evidence on Harms of Medical Interventions in Randomized and Nonrandomized Studies." *Canadian Medical Association Journal* 174 (5): 635–641. doi: 10.1503/cmaj.050873.

Papanikolaou, P. N., R. Churchill, K. Wahlbeck, and J. P. Ioannidis. 2004. "Safety reporting in randomized trials of mental health interventions." *American Journal of Psychiatry* 161 (9): 1692–1697. doi: 10.1176/appi.ajp.161.9.1692.

Pitrou, I., I. Boutron, N. Ahmad, N., and P. Ravaud. 2009. "Reporting Of Safety Results in Published Reports of Randomized Controlled

Trials." *Archives of Internal Medicine* 169 (19): 1756–1761. doi: 10.1001/archinternmed.2009.306.

Rawlins, M. 2008. *De Testimonio: On the Evidence for Decisions about the Use of Therapeutic Interventions*. London: Royal College of Physicians. http://www.amcp.org/WorkArea/DownloadAsset.aspx?id=12451 (accessed 9 November 2015)

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

Stegenga, J. 2011. "Is Meta-Analysis the Platinum Standard?" *Studies in History and Philosophy of Biology and Biomedical Sciences* 42: 497–507.

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

Stegenga, J. (forthcoming-c). *Medical Nihilism*. Oxford: Oxford University Press.

Straus, S. E., Richardson, W. S., Glasziou, P. P., & Haynes, R. B. 2005. *Evidence-based Medicine: How to Practice and Teach It*, 3rd ed. London: Elsevier Churchill Livingstone.

Swanson, J. M., G. R. Elliott, L. L. Greenhill, T. Wigal, L. E. Arnold, B. Vitiello, L. Hechtman, J. N. Epstein, W. E. Pelham, H. B. Abikoff, J. H. Newcorn, B. S. Molina, S. P. Hinshaw, K. C. Wells, B. Hoza, P. S. Jensen, R. D. Gibbons, K. Hur, A. Stehli, M. Davies, J. S. March, C. K. Conners, M. Caron, N. D. Volkow. 2007. "Effects of Stimulant Medication on Growth Rates across 3 Years in the MTA Follow-up." *Journal of the American Academy of Child and Adolescent Psychiatry* 46 (8): 1015–1027. doi: 10.1097/chi.0b013e3180686d7e.

Tsang, R., Colley, L., and L. D. Lynd. 2009. "Inadequate Statistical Power to Detect Clinically Significant Differences in Adverse Event Rates in Randomized Controlled Trials." *Journal of Clinical Epidemiology* 62 (6): 609–616.

Vandenbroucke, J. P. 2008. "Observational Research, Randomised Trials, and Two Views of Medical Science." *PLoS Med* 5 (3): e67. doi: 10.1371/journal.pmed.0050067.

Whitaker, R. 2010. *Anatomy of an Epidemic: Magic Bullets, Psychiatric Drugs, and the Astonishing Rise of Mental Illness in America*. Danvers, MA: Crown.

Worrall, J. 2002. "What Evidence in Evidence-Based Medicine?" *Philosophy of Science* 69: S316–S330.

Worrall, J. 2010. "Do We Need Some Large, Simple Randomized Trials in Medicine?" Pp. 289–301 in *EPSA Philosophical Issues in the Sciences*. Edited by M. Suarez, M. Dorato, and M. Redei. Dordrecht: Springer.