Robust and Discordant Evidence: Methodological Lessons from Clinical Research

Spencer Phillips Hey*†

The concordance of results that are "robust" across multiple scientific modalities is widely considered to play a critical role in the epistemology of science. But what should we make of those cases where such multimodal evidence is discordant? Jacob Stegenga has recently argued that robustness is "worse than useless" in these cases, suggesting that "different kinds of evidence cannot be combined in a coherent way." In this article I respond to this critique and illustrate the critical methodological role that robustness plays as an aim of scientific inquiry.

1. Introduction. Concordant evidence from multiple investigations is widely considered to play a critical role in the epistemology of science. When multiple experiments agree on a result or when multiple models can derive the same theorem, this is usually taken as confirmation—providing good grounds for thinking that the common result is of genuine interest and not merely an artifact of some simplifying assumption. More generally, the epistemic virtues of concordant evidence fall under the philosophical category of "robustness." As described in Levins (1966), Wimsatt (1981, 2007), and numerous essays in the recent Soler et al. (2012) collection, both robustness, a property of hypotheses when they are supported by evidence from multiple modalities (hereafter just "multimodal evidence"), and robustness analysis, a systematic search for robust hypotheses, are taken as methodologically fundamental to many scientific activities. This includes distin-

Received January 2014; revised April 2014.

*To contact the author, please write to: McGill University, 3647 Peel Street, Montreal, QC H3A 1X1, Canada; e-mail: spencer.hey@mcgill.ca.

†I would like to thank Charles Weijer, Robert Batterman, Charles Heilig, William Wimsatt, and the anonymous reviewers of this journal for their insightful feedback on this manuscript.

Philosophy of Science, 82 (January 2015) pp. 55–75. 0031-8248/2015/8201-0007$10.00

guishing between reliable and unreliable evidence, true and false theories, useful and biased assumptions, and so on.

But what should we make of those cases where multimodal evidence is discordant? Wimsatt (1980) discusses one such case involving discordance between mathematical and laboratory models of group selection in the late 1960s to early 1980s. Franklin (2002) discusses another involving alternative interpretations of the Liquid Scintillator Neutrino Detector experiments in physics in the early 1990s. Both of these authors see robustness and robustness analysis as key to resolving the discordance. However, Stegenga (2009, 2012) has recently argued that these philosophical responses are not true solutions to the problem of multimodal discordance.[1] Instead, he claims, they merely point to the sources of the problem:

> If different modes of evidence support contrary conclusions, there is no obvious way to compare or combine such evidence in an orderly or quantifiable way, let alone to compare such a combination of evidence to evidence from a single mode. Philosophers have long wished to quantify the degree of support that evidence provides to a hypothesis. At best, the problem of discordance suggests that robustness is limited to a qualitative notion. And if robustness is a qualitative notion, how should we demarcate robust from non-robust evidence? At worst, the problem of discordance suggests that evidence of different kinds cannot be combined in a coherent way. (Stegenga 2012, 215)

Although other critics of robustness have introduced counterexamples that seem to illustrate a failure of robustness analysis,[2] Stegenga's critique is the first sustained philosophical argument (to my knowledge) against the fundamental epistemic value of robustness. In this essay, I argue that his critique is significantly flawed. Nevertheless, discussion of his argument is instructive for pointing us toward a critical lacuna in the philosophical discussion of robustness—namely, the need to more precisely articulate the function of robustness analysis in resolving discordant multimodal evidence.

I begin in the next section by reviewing Stegenga's argument that "robustness-style arguments do not tell us what to believe in situations of evidential discordance" (2012, 216). Then, in section 3, I offer an immediate reply, which largely blunts the force of his critique. However, in section 4, I present a case study drawn from the domain of clinical trials and translational medicine, which more accurately elucidates the epistemolog-

1. The ideas in Stegenga (2009) and those in Stegenga (2012) on discordant evidence are nearly identical, but for the purposes of this essay I will focus solely on the more recent publication.

2. See, e.g., Orzack and Sober (1993) and the reply from Levins (1993), or Rasmussen (1993) and the reply from Culp (1994).

ical complications and challenges arising from multimodal discordance. In section 5, I show how these challenges can be overcome with robustness arguments—specifically, I illustrate the critical methodological role that robustness considerations play in generating new explanatory hypotheses and driving future research forward. I ultimately conclude by reemphasizing the importance of the distinction between the property of robustness, which may at times be elusive in the face of discordant multimodal evidence, and the strategy of robustness analysis, which remains a vital component of scientific methodology and epistemology.

2. Stegenga on Discordance. Stegenga's (2012) argument can be broken down into two claims: (1) since multimodal evidence is usually discordant, the defenders of robustness need to elucidate what role (if any) robustness has to play in resolving the discordance; and (2) in the absence of systematic or principled methods for amalgamating different modes of evidence, robustness arguments are vague and provide little epistemic guidance.[3]

In support of claim 1, Stegenga provides us the example of influenza transmission. As he describes it, there is controversy in the medical community as to whether influenza is transmitted through contact or through the air. Physicians and nurses tend to believe that it is spread through contact, whereas academic virologists tend to believe that it is airborne. Complex mathematical models seem to suggest that the probability of airborne transmission is high; however, the evidence from animal and controlled human experiments has been equivocal. Since these different modalities produce conflicting results and there is "no obvious way to compare or combine" the evidence, Stegenga claims that expert opinion about the true mode of transmission remains divided.

Stegenga takes this case to demonstrate "the poverty of robustness" (2012, 215). The proponents of robustness have insisted that multimodal evidence is supposed to increase the reliability of results, but when the evidence is discordant, it appears to do just the opposite. As he puts it, "In such cases robustness is worse than useless, since the fact of multiple modes of evidence is the source of the problem" (216).

Of course, as Stegenga acknowledges, one example does not prove that multimodal evidence is always so problematic. Nor does it prove that evidence is more often discordant than concordant. Nonetheless, he asserts that

3. Stegenga also discusses some other issues related to robustness arguments, including (*a*) the problem of needing multiple investigational modalities when there may not be any, (*b*) the problem of determining whether or how the multiple modalities are sufficiently independent, and (*c*) the distinction between security (as described in Staley 2004) and robustness. Although these are also interesting philosophical problems, since they are separable from the problem of multimodal discordance, I will not discuss them here.

this type of scenario is the more common situation in science, writing that "history of science might occasionally provide examples of apparent concordance, but concordance is easier to see in retrospect, with a selective filter for reconstructions of scientific success" (2012, 216). In other words, the proponents of robustness have disproportionately emphasized those fortuitous instances of concordant multimodal evidence from the history of science. So even if robustness is a useful concept when multimodal evidence is concordant, the rarity of this situation significantly mitigates the epistemic value of robustness.

In support of claim 2—that robustness arguments are vague—Stegenga begins by providing a useful distinction between two dimensions of discordance: inconsistency and incongruity. Inconsistency is contradictory evidence among similar types of investigations: one animal experiment shows x, and the next animal experiment shows not-x. Incongruity is (apparently) incommensurable evidence among different types of investigations: a mathematical model shows x, an animal model shows y, an epidemiological study shows $0.5y$, a clinical trial shows $2x$, and clinical experience shows z. While incongruous results do not all clearly contradict one another (as is the case for inconsistency), it is still unclear how one should assemble a coherent picture of the total evidence.

Indeed, for both inconsistency and incongruity, Stegenga argues that there is no universal method to amalgamate the evidence. Since the evidence from these different scientific modalities is written in different "languages," the background assumptions necessary to "translate" the results from one language to the next "will have varying degrees of plausibility" (2012, 214). Even for instances of concordant evidence, if the assumptions necessary to "translate" the evidence are not plausible, then it is not clear how different investigational modalities (e.g., mathematical models vs. animal experiments vs. human experiments) are supposed to support a robustness claim.

For discordant evidence, the problem is further complicated by the necessity of sorting out the overall direction of the evidence. Which of the many hypotheses are actually supported by the total evidence? Or which hypotheses should be revised? As Stegenga puts it, "In the absence of a methodological meta-standard, there is no obvious way to reconcile various kinds of inconsistent data" (2012, 213). And if we cannot reconcile the data, then, again, this seems to undermine the epistemic value of robustness arguments.

Ultimately, Stegenga argues that the answer to these problems lies in articulating better "amalgamation functions." Such functions would tell us how to gather all the pieces of evidence, weight them according to various epistemic properties (e.g., quality, relevance, salience, and concordance), and then output the degree of support provided (or the correct credence) for a given hypothesis. He then mentions a number of possible functions, some

of which take quantitative input from multiple modes and provide a quantitative output, such as statistical meta-analysis, and others that take quantitative input but provide qualitative output, such as the evidence-based medicine hierarchy (2012, 222). He also considers the possibility of a Bayesian amalgamation function built on Jeffrey conditionalization (222–24). But in the end, he finds all of these functions wanting and concludes that "without the use of compelling schemes to amalgamate discordant multimodal evidence, robustness arguments are vague" (222).

3. Reply. There is much in Stegenga's argument with which I agree. I would agree that multimodal discordance is probably the more common scientific reality. I agree that inconsistent evidence and incongruent evidence present interpretive challenges for researchers. And I agree that better functions for amalgamating evidence would be a boon for the epistemology of science. Yet, his argument is fundamentally flawed in a number of ways.

To begin with, much of Stegenga's argument relies on a false dichotomy between "principled and systematic" methods of evidence assessment, which we are to understand as rigorous and quantitative, and the "merely intuitive or qualitative" methods, which he insists must be "disorderly" or "vague." Consider, for example, his rhetorical question: "And if robustness is a qualitative notion, how should we demarcate robust from nonrobust evidence?" This is immediately followed by the suggestion that if robustness is qualitative, then "evidence of different kinds cannot be combined in a coherent way" (2012, 215).

On the contrary, there can be principled, systematic, and yet qualitative methods of evidence evaluation. He even acknowledges one such method—systematic reviews—at the conclusion of his argument. That such qualitative methods are not foolproof does not entail that they cannot still be rigorous, valid, or useful for amalgamating evidence.

Similarly, there is no a priori reason to expect that robustness analysis should be "obvious" or straightforward to apply. Therefore, Stegenga's frequent objection that there is "no obvious way" to combine evidence from different modes of evidence is not well motivated. The seminal philosophical discussions of robustness analysis consistently acknowledge that there are limitations, complications, and conditions on its valid use (Wimsatt 1980, 1981; Levins 1993). Therefore, in order to support his conclusion that robustness arguments are too vague and problematic, he would need to show that these conditions for valid application of the concept are routinely ignored, that they cannot be satisfied even in principle, or that they are poorly specified.

I take it that the influenza transmission example is supposed to do some of this work. Stegenga claims that this example illustrates the "poverty of

robustness" because, despite there being evidence from multiple scientific modalities, the expert medical community remains uncertain about the mode of transmission. But for this example to be compelling, it is necessary that airborne and contact transmission vectors are mutually exclusive—and that claim is, at best, controversial (if not simply false). The possibility of multiple disease transmission mechanisms has been recognized by the medical and public health communities for over 100 years (Chapin 1910). Moreover, both the US and European centers for disease control explicitly state that influenza can be spread in either fashion.[4] And if these transmission vectors are not mutually exclusive, then different research modalities providing evidence for the effects of the different vectors are not a problem for robustness at all. The multiple modalities can be understood as modeling different (albeit related) systems.[5]

If we can thus reject the influenza case as an exemplar of multimodal discordance, then this undercuts the philosophical force of Stegenga's assertions that discordant evidence is the more common occurrence. He states that he is not arguing against the common methodological prescription for scientists to "get more data" when they encounter conflicting evidence from multiple modalities. Rather, he is only arguing that "robustness-style arguments do not tell us what to believe in situations of evidential discordance" (2012, 216). But the proponents of robustness could now simply point to the influenza example as showing that even if discordant multimodal evidence is the more common epistemic state in science, so long as scientists adopt the methodological prescriptions of robustness analysis (e.g., "get more data, and of different kinds!") in the face of discordant evidence, they are thereby able to sufficiently resolve the uncertainty. Contrary to robustness being "worse than useless," the influenza example now makes it appear as if robustness is working exactly as it should. Perhaps there are no universal amalgamation functions, but the practical necessity of these depends on multimodal discordance being epistemologically vexing, and Stegenga has failed to show that it is.

4. See http://www.cdc.gov/flu/about/disease/spread.htm (retrieved November 19, 2013) and http://ecdc.europa.eu/en/healthtopics/Documents/#0905 (retrieved November 19, 2013).

5. It is difficult to know what to make of Stegenga's conclusions here, since he does not provide any references for the supposed discordant models and experiments concerning influenza transmission. But from my reading of the literature, the existence of multiple disease vectors seems to be the consensus (Weber and Stilianakis 2008). Further, as Hall (2007) describes it, the scientific controversy is not about the mode of infection—since the spread of influenza is known to vary from case to case (or even person to person) and known to depend on a range of different physical (e.g., aerosol droplet size) and social conditions (e.g., compliance rates for hand-washing regimes among health care personnel); rather, the lingering scientific uncertainty is about the most effective means to intervene and reduce the spread of disease.

On the other hand, even if we were to accept Stegenga's claims that genuine multimodal discordance and epistemic controversy are the norm, this is only problematic for the property of robustness. That is, it may indeed be challenging to resolve cases of discordance in order to definitively conclude that a particular hypothesis is robust; however, this does not entail that systematic investigations to try and resolve such states of discordance are necessarily futile. This underscores the importance of distinguishing between robustness as a property of particular hypotheses and robustness as an aim of inquiry. The epistemic goal of a robust hypothesis may prove elusive, but this does not mean that we ought to abandon that goal.

4. Discordance across a Clinical Research Trajectory. These criticisms notwithstanding, I think it would be overhasty to entirely dismiss Stegenga's concerns. As I suggested at the outset, there is still an important philosophical problem here. But to better illustrate the challenges that discordance presents for robustness, it will be helpful to focus on a genuine instance of multimodal discordance.

To that end, let us consider the structure of clinical research: In the ideal sequence of testing for any new medical intervention, a consistent and congruent pattern of efficacy and effectiveness is observed across a series of experimental modalities (typically called "phases")—from the in vitro and in vivo preclinical experiments to the phase 1, 2, and 3 human clinical trials. Each successive, positive outcome in these experiments increases our confidence in the reality of the treatment's net therapeutic advantage, and this growing confidence in turn discharges the ethical demands of clinical equipoise, that is, that there exist, before initiating the pivotal phase 3 trials, a state of honest, professional uncertainty among the expert medical community as to the therapeutic merits of each arm in the study (Freedman 1987; Weijer and Miller 2004).

We can immediately observe that robustness arguments are integrated into the very structure of this research model. That is, an effective medical intervention is a robust intervention—it has been tested in a dish, in an animal, in healthy volunteers, and in patients. Moreover, despite the apparent hierarchical structure of these phases, we should also observe that clinical translation is not a linear or monotonic process. Phases will often overlap with each other and may be repeated or iterated in light of the accumulating state of total evidence.

For example, Mateo et al. (2013) recently described the sequence of testing for the anticancer agent iniparib. This drug was initially developed as a poly(ADP-ribose) polymerase (PARP) inhibitor—a class of drugs whose mechanism of action impairs DNA repair functions (and hence kills the cancer cells). Based on this mechanistic theory, iniparib was tested on

over 2,500 patients across all three phases of clinical trials. However, after its failure in phase 3, iniparib was retested in vitro, where its status as a genuine PARP inhibitor was called into question. This shows how negative results in the later phases of clinical testing may suggest a necessary revision of the driving pathophysiological theories. These new theoretical hypotheses may then be tested in preclinical experiments and, if confirmed, human trials reinitiated.[6]

Just as Stegenga would have it, the iniparib case is far from exceptional. Discordant results are common across the trajectory of clinical translation (Hay et al. 2014). Clinical investigators are regularly confronting discordant multimodal evidence when deciding how best to proceed. A recent series of experiments testing the efficacy of the antibacterial agent moxifloxacin for treating tuberculosis provides another useful illustration of this problem: after a concordant and positive trend through the preclinical and early clinical phases, five phase 2 trials produced discordant evidence—two trials were negative, three trials were positive—and it was unclear how or whether research with moxifloxacin ought to proceed.

In contrast to the iniparib case, whose theory has already been put to decisive testing in phase 3 trials, with moxifloxacin we are concerned with whether or not more decisive testing is warranted. I present the details of these experiments in the remainder of this section, drawing attention to the function and utility of robustness considerations throughout. As I will show, despite multimodal discordance at nearly every stage of the research process, the scientists were nevertheless able to construct valid and justified robustness arguments to drive the research program forward.

4.1. Moxifloxacin Trials. For drug-susceptible tuberculosis, the current standard treatment is a four-drug regimen (isoniazid, rifampin, pyrazinamide, and ethambutol) administered for 6 months. Since what ultimately matters is not just eliminating the active bacteria (which can happen quickly) but preventing future relapse, a large part of the challenge to eliminating tuberculosis is ensuring full compliance with this regimen for 6 months. Thus, the central aim of new research is to find a shorter or simpler regimen. Unfortunately, evaluating the clinical outcome of "no future relapse" is costly, requiring at least 2 years of patient follow-up. Such an undertaking is typically not possible until the phase 3 stage of research. As a result, most of the earlier efficacy studies for new tuberculosis treatments adopt 8-week culture conversion as a surrogate endpoint. This means that a patient whose sputum culture has converted to tuberculosis-negative within

6. See also Hey, Heilig, and Weijer (2013), who introduce a graph-theoretic model for representing the complex interactions among the various experimental modalities in clinical research.

8 weeks of beginning treatment is considered a positive outcome.[7] Earlier-stage studies adopt various other surrogate endpoints, depending on the model in question.

However, the use of these surrogate endpoints in the early phases—which is standard practice throughout most domains of clinical research—generates an interpretive problem for researchers as they approach the so-called go/no-go decision at the cusp of phase 3 testing. Phase 3 trials are the last stage before a new drug is submitted to the national regulatory agencies for approval. These are often large, expensive, and time-consuming experiments, representing an enormous material investment for both the scientists and the research funders. It is therefore of critical importance that earlier trials use predictive surrogates—that is, experimental outcomes that can be assessed more quickly but are still well correlated with the clinical outcome.

This relationship between surrogate and clinical outcome necessarily mediates the strength of evidence between phases. A converted sputum sample does not necessarily mean that the patient is cured of tuberculosis. Nor does an effective cure in mice mean that the same drug will be effective in humans. Thus, at every phase transition—from preclinical to clinical, from phase 1 to phase 2, and from phase 2 to phase 3—an inference must be made about the state of accumulating evidence and the justification for initiating the next phase.

For example, when initiating clinical trials, investigators will want to see evidence of robustness across the in vitro (e.g., agent kills tuberculosis bacteria) and animal modalities (e.g., agent reduces tuberculosis colonies in the lungs of different strains of mice). When considering phase 2 trials, investigators want to see a continuing trend of robust evidence through the phase 1 modality (e.g., early bactericidal activity and acceptable toxicity), as well as any of the continuing animal or in vitro investigations.

I should also emphasize that the relevant body of evidence is not simply the most recent evidence produced by the current (or immediately preceding) phase. Although phase 1 evidence is, in some respects, the most directly relevant when contemplating phase 2 (and mutatis mutandis for the other phase transitions), it is not the case that the evidence from each later phase "trumps" what has come before. It truly is an accumulating body of multimodal evidence that supports decision making at every step

7. Sputum is material expelled from the patient's lungs or collected from their saliva. If a patient has a bacterial infection, then this material will contain the microbacteria of interest. The sputum sample is collected (and possibly stored) by the researchers, so that the bacteria can be cultivated in a growth medium, either a solid agar medium (sometimes called a "plate") or a liquid broth medium. The amount of viable bacteria found in the sputum, usually measured in cfu/ml (colony forming units per milliliter), is a surrogate for the amount of bacteria in the patient's lungs.

along a clinical translation. In fact, as we will shortly see with moxifloxacin, evidence from ongoing preclinical trials played a critical role in shaping the hypothesis that was pursued in a later phase 2 trial.

Let us now turn to the details: The first experiments testing moxifloxacin against tuberculosis were the two in vitro studies, Ji et al. (1998) and Gillespie and Billington (1999). Both of these experiments tested a range of agents against tuberculosis bacteria and agreed that moxifloxacin was the most promising. Gillespie and Billington (1999) also noted that moxifloxacin's in vitro properties were similar to isoniazid, one of the sterilizing drugs already in the standard antituberculosis treatment regimen. These studies were followed by Miyazaki et al. (1999), the first in vivo mouse experiment, which confirmed moxifloxacin's similar performance to isoniazid—a congruence between in vitro and in vivo studies—and argued that this was a sign of positive potential for its use in multidrug treatment of human tuberculosis.

The next mouse experiment was Lounis et al. (2001), who observed that the addition of moxifloxacin to a 6-month weekly dosage regimen was also only marginally inferior to the 6-month daily dosage standard, cautiously concluding that moxifloxacin might have potential for simplifying the standard regimen. However, shortly thereafter Yoshimatsu et al. (2002) published a discordant result: they did not see any significant bactericidal effect from the weekly dosages of moxifloxacin in their mouse experiment. Given that Yoshimatsu et al.'s was a test of moxifloxacin as monotherapy, they suggest that Lounis et al.'s earlier result could have been due to interaction effects with moxifloxacin and the other drugs in the regimen. They also note some "untoward" side effects from higher dosages of moxifloxacin: failure to gain weight, decreased activity, and unkempt fur after 4 weeks. Thus, they concluded that high daily dosages of moxifloxacin were promising, but further studies of weekly combination therapies with moxifloxacin, as well as its toxicity effects on healthy mice, were still needed.

Gosling et al. (2003) was the first phase 1, early bactericidal activity (EBA) test of moxifloxacin. Their results were concordant with the in vitro and in vivo results of moxifloxacin's bactericidal activity, again finding moxifloxacin to be similar in activity to isoniazid. Given this robust trend of similarity to isoniazid, they felt ready to conclude that "clinical trials to determine whether regimens containing moxifloxacin bring higher rates of culture conversion at two months should be performed as soon as sufficient safety data are available" (Gosling et al. 2003, 1345).

This recommendation was largely shared by Nuermberger et al. (2004), in their subsequently published mouse experiment. However, Nuermberger et al. did not see a significant reduction in culture conversion time. Instead, they observed a "dramatic increase in potency" when moxifloxacin was substituted for isoniazid in the standard treatment regime (424). Since

reduction in overall treatment time is one of the major goals of tuberculosis research, this result is, at best, mixed. Nevertheless, their articulation of a novel hypothesis—substituting moxifloxacin for isoniazid in the standard regimen—is an important development, since this hypothesis directly informs the design of the Dorman et al. (2009) phase 2 study to be discussed below.

Three more phase 1 EBA studies followed: Pletz et al. (2004) was largely consistent with Gosling et al.'s early finding and showed moxifloxacin to be very similar to isoniazid. Gillespie et al. (2005) produced an inconsistent result, finding no significant improvement with the combination of moxifloxacin and isoniazid (although this is arguably concordant with Nuermberger et al. 2004). Johnson et al. (2006) again showed minor improvements in EBA with moxifloxacin.

Before discussing the phase 2 experiments, let us pause here and consider the role that robustness considerations played across the preclinical and phase 1 studies. Indeed, there was already some evidence of discordance. For example, Lounis et al. (2001) and Yoshimatsu et al. (2002) were inconsistent on whether a simpler, weekly dosage regimen (as opposed to the standard daily regimen) with moxifloxacin showed acceptable in vivo efficacy. The unacceptable toxicity seen in Yoshimatsu et al. (2002) and the lack of EBA in Gillespie et al. (2005) are also inconsistent with other similar studies.

But despite the multimodal discordance, there is still a robustness argument in support of the decision to initiate phase 2 trials: first, moxifloxacin was already known to be effective across a range of other antibacterial indications;[8] second, it was consistently shown to be well tolerated and acceptably safe in humans (Stass et al. 1998); third, the in vitro and in vivo studies consistently and concordantly showed that moxifloxacin had activity against tuberculosis; and fourth, its antituberculosis activity was consistently and concordantly shown to be similar to that of isoniazid—one of the drugs already in the standard regimen. Indeed, the expert community was largely in agreement—even before 2004—that the evidence of moxifloxacin's safety and efficacy was robust enough to warrant phase 2 trials.[9]

8. The fluoroquinolone family of drugs is a widely used and extensively tested class of antimicrobial agents, so much so that moxifloxacin, a single member of this class, warranted its own supplement in the journal *Clinical Infectious Diseases* in 2005—just 5 years after it was first approved for testing. In addition to analyses of the pharmacokinetics, pharmacodynamics, and safety, the articles in that volume discuss moxifloxacin's effectiveness for treating pneumonia, rhinosinusitis, and acute exacerbations of chronic obstructive pulmonary disease.

9. Nuermberger et al. (2004) explicitly note that their "experimental results support the rationale for *ongoing randomized clinical trials* designed to test whether the addition of [moxifloxacin] to the standard regimen will increase the proportion of patients with

Yet, it is important to distinguish the biological, methodological, and practical questions at issue here. The above argument supports an affirmative answer to the biological question "Is moxifloxacin active against tuberculosis bacteria?" It also arguably supports an affirmative answer to the methodological question "Is a phase 2 trial justified?" But it does not support an affirmative answer to the practical (or clinical) question "Is moxifloxacin an effective treatment for tuberculosis?" This question is the responsibility of the phase 2 and 3 trials.

Burman et al. (2006) was the first published phase 2 study. They analyzed 277 patients across multiple sites in the United States and Africa, randomized to either (1) the standard four-drug regimen or (2) the standard with moxifloxacin substituted for ethambutol.[10] Sputum cultures were grown and analyzed using both liquid and solid media. Although moxifloxacin showed possible increased activity at earlier time points, it did not affect 8-week sputum culture status. They also stratified their analysis by continent and found that African patients, despite the highest rate of compliance, responded far less to either treatment than did patients in the United States. They concluded that further research with moxifloxacin was needed, but it seemed unlikely to shorten the overall treatment time for tuberculosis.

Rustomjee et al. (2008) was the next phase 2 study published. Theirs was a four-arm study comparing (1) the standard regimen versus three different flouroquinolone substitutions for ethambutol: (2) ofloxacin, (3) gatifloxacin, and (4) moxifloxacin. It was conducted at a single site in Durban, South Africa, and, like Burman et al. (2006), evaluated sputum cultures with both liquid and solid media. However, unlike Burman et al., they did not use the 8-week culture conversion status as the surrogate endpoint, instead adopting the rate at which cultures converted to tuberculosis-negative.

They analyzed 217 patients in total (approximately 55 patients per arm) and found that both gatifloxacin and moxifloxacin improved the rate at which sputum cultures converted to tuberculosis-negative. As in Burman et al.'s study before them, neither moxifloxacin nor gatifloxacin showed any increased effect on 8-week sputum culture status. Nevertheless, because of their alternative endpoint, Rustomjee et al. take their result to support the opposite conclusion. They also note that a significant difference between the moxifloxacin and control arms was only found with cultures grown in solid media. Whether despite or because of these breaks in methodology

---

negative sputum cultures after two months of therapy" (2004, 424; emphasis added). Although it is unclear exactly to which studies they are referring, at the very least, Burman et al. (2006) would have been under way at this time.

10. Ethambutol is most often used as the comparator drug in the phase 2 trials since it is thought to be the weakest drug in the current regimen (Steenwinkel et al. 2010).

with Burman et al.'s previous work, they nevertheless conclude that a phase 3 trial with moxifloxacin is warranted.

Conde et al. (2009) was another single-site (Rio de Janeiro, Brazil), two-arm study comparing the standard regimen to the substitution of moxifloxacin for ethambutol. Like Burman et al., they used the 8-week conversion as the surrogate endpoint, but unlike the two earlier studies, they used only solid media cultures. Analyzing a total of 125 patients, they found a significant difference favoring the moxifloxacin arm.

Dorman et al. (2009) was another multisite study, this time with locations in South Africa, Uganda, North America, Brazil, and Spain, which investigated the standard regimen versus a regimen where moxifloxacin was substituted for isoniazid. As I noted, this hypothesis was justified on the basis of results from mouse models (Nuermberger et al. 2004). Dorman et al. used 2-month culture conversion status as their surrogate endpoint and evaluated cultures of both liquid and solid media.

After analyzing 328 patients, 213 of whom came from African sites (65%), they found no significant difference between the moxifloxacin and control arms. They were able to show that enrollment at an African site was associated with a lower likelihood of 8-week culture conversion status (regardless of the treatment received) but could not conclude that the prospects for reduced treatment time with moxifloxacin were positive.

Wang et al. (2009) was an open-label, single-arm study conducted at a single site in Taiwan. Rather than a substitution, they compared the standard four-drug regimen with the standard plus moxifloxacin. Cultures were analyzed on both liquid and solid media, but unlike any of the previous studies, the surrogate endpoint adopted was 6-week (rather than the usual 8-week) culture conversion, on the grounds that "using [2-month culture conversion] alone as the primary endpoint does not reflect the entire spectrum of effectiveness of a fluoroquinolone-containing anti-tuberculosis regimen" (Wang et al. 2009, 65). They analyzed 123 patients and found a significant improvement with the moxifloxacin arm at 6 weeks.

As of 2010, this was the accumulated state of evidence for tuberculosis researchers—after a promising trend in preclinical and phase 1 trials, there were three positive phase 2 studies and two negative or null phase 2 studies. So what should the tuberculosis researchers believe? How should research on moxifloxacin proceed? We were able to identify a robustness argument to support initiation of phase 2, but is there an equivalent argument available here?

4.2. Revising the Problem of Discordance. Before discussing how robustness analysis contributes to the resolution of this problem, there are a number of philosophical points to emphasize. First, this scenario illustrates a further complication that multimodal discordance poses for robustness.

That is, even if we assumed, along with the proponents of robustness, that over the long term a concordant result will emerge from further multimodal investigations, this still does not mitigate the more immediate epistemological concern about the justifiable beliefs and research strategies in the midst of a given state of discordant evidence. In other words, what should the scientist believe when she is confronting discordant evidence? She cannot (or may not want to) simply wait for the passing of time and additional experiments to sort it out. Even if she can obey the prescription to "gather more data," this still does not tell her what kinds of data to gather or which investigational modalities would be the most epistemically valuable. It also does not tell her when she might be better served by abandoning her present line of inquiry and investing in an entirely different approach.

In essence, this challenge combines the familiar epistemic concerns arising from underdetermination with some additional dimensions of methodological uncertainty due to multimodal discordance. In so doing, it more effectively shifts the burden of proof onto the philosophical defenders of robustness. That is, however illuminating may be the rational reconstructions or historical analyses for which multimodal concordance eventually emerged, they do not tell us how the scientists may have (or should have) reasoned in the midst of discordance. Nor do they tell us what role (if any) robustness considerations had to play in resolving the discordance (beyond merely exhorting the scientists to gather more data). If robustness and robustness analysis are as methodologically important as its defenders claim, then they need to show how it can contribute to the systematic amalgamation of discordant evidence in real time. And this is precisely what is at issue for tuberculosis researchers: Is the evidence for moxifloxacin's effectiveness robust enough to warrant phase 3 trials? If not, then what studies ought to be done next?

Second, we should observe that in medical research, particularly at the stage of clinical trials, cost can be a serious constraint on the possible number of further investigations. Not only is there a material cost to designing and conducting a trial, but there is also the opportunity cost—the time lost by investigators and research subjects in testing one course of treatment rather than another. Even a well-funded research program, capable of running many early-phase studies, might be better off cutting its losses and abandoning an unpromising experimental treatment in order to pursue something else. This adds additional weight to the decision facing the researchers and calls into question the value of the standard prescription from robustness analysis to simply gather more data until the discordance is resolved.

Third, it is tempting to appeal to the use of statistical meta-analysis as a means for resolving the question. On its face, the trials show no clear trend toward effectiveness or lack thereof, but perhaps by pooling together the data provided by each study, more decisive evidence could be detected.

However, a valid statistical meta-analysis critically depends on the homogeneity of the studies and their data. Studies with different populations, using different outcome measures, or with different measurement techniques are much harder (if not impossible) to validly pool together. While this kind of heterogeneity between studies does not categorically rule out the usefulness of a meta-analysis, its validity becomes dependent on arguments for why the heterogeneity can be permitted.

In the case of the moxifloxacin trials, although there is some evidence of homogeneity across the studies, particularly in the measurement techniques (as discussed in the methods section of each publication), the discussion sections across the studies reveal an underlying heterogeneity in methodology. For example, Rustomjee et al. (2008) are critical of the binary outcome measure, "culture negative at some time *t*," and question whether or not this is the appropriate surrogate endpoint. Insofar as their critique is well founded, it significantly complicates the structure of any meta-analysis, since "culture negative at some time *t*" is the endpoint adopted by four of the other studies. Employing a meta-analysis with their preferred alternative endpoint (i.e., rate of conversion to tuberculosis-negative) would require that the data from the four other studies be reanalyzed. Since both Burman et al. (2006) and Dorman et al. (2009) sampled the culture status at intervals of 2 weeks (half as often as Rustomjee et al. 2008), the requisite data may not even be available.

Rustomjee et al. are also critical of using liquid media to grow and analyze the collected sputum cultures. Since the liquid media results of their study showed no significant difference between moxifloxacin and the control at 8 weeks, they argue that solid media and the continuous endpoint of rate of conversion "may be a more useful method of assessment" (Rustomjee et al. 2008, 135). But such a meta-analysis, excluding the liquid media results from all of the studies (excepting Conde et al. [2009], who did not use liquid media), would effectively beg the question. To test Rustomjee et al.'s suggestion with the extant data, we would have to assume that moxifloxacin is effective and then go back to see whether liquid media across the studies failed to indicate its effectiveness. This runs the risk of finding a significant result simply through data mining.

Finally, there are questions about the internal validity of the studies. In both Burman et al. (2006) and Dorman et al. (2009), there was a significant difference in treatment response between the patients in Africa and the patients elsewhere in the world. As statistical outliers, there is a temptation to exclude them from a meta-analysis. Perhaps African patients represent a relevant treatment subgroup that does not respond as well to moxifloxacin and thus should be excluded to avoid biasing moxifloxacin's effectiveness in the overall drug-susceptible tuberculosis population.

It is an important question as to why this difference was observed in African patients in the two null studies. But to exclude them in a meta-

analysis again begs the question in favor of moxifloxacin's effectiveness. Dorman et al. (2009) are right to conclude cautiously that treatment at an African site in their study was "associated" with a worse outcome. It could very well be that a medically relevant subpopulation for treating tuberculosis has been identified. More to the point, however, is the fact that tuberculosis is most prevalent in Africa: 30% of all new cases of tuberculosis are in Africa, as well as 80% of tuberculosis–human immunodeficiency virus (HIV) coinfection. For the global effort toward treating and controlling tuberculosis, a treatment that does not work on African patients is ultimately of very little interest.

This is all to argue that the heterogeneity across these studies precludes the usefulness of a statistical meta-analysis, and therefore this possible resolution to the discordance is not available.

5. The Function of Multimodal Robustness. Thus, we are still left with some questions: How should tuberculosis researchers respond to the state of discordance? Is the evidence of moxifloxacin's efficacy against tuberculosis sufficient to initiate phase 3 trials or not? If the simple methodological prescription of robustness to conduct more phase 2 trials is imprudent or cost-prohibitive and the utility of a statistical meta-analysis is questionable, what other strategies are available?

While some evaluation techniques, such as statistical meta-analysis, are weakened by experimental heterogeneity, robustness analysis is actually strengthened. Indeed, as Wimsatt has long argued, contrasting the properties of experiments or models where robustness fails can be informative about the system features on which a result critically depends (Wimsatt 2007). Robustness analysis—driven by the aim of achieving a robust result—thus helps to answer the immediate and pressing concern about what kind of evidence is needed.

For example, the discordance in phase 2 motivated a retrospective study by Mac Kenzie et al. (2011), which showed that the observed difference between African and non-African outcomes in Dorman et al.'s (2009) study is not explained by baseline severity of the disease, HIV status, age, smoking, diabetes, or race. Similarly, a subsequent in vitro experiment from Shandil et al. (2007) explored whether drug interaction effects within the moxifloxacin regimens could explain the observed discordance between the promising animal results and the disappointing human results.

These follow-up investigations, searching to explain the failures of robustness (in a relatively inexpensive experimental setting), elucidate one of the key philosophical points: there is much more going on across the chain of multimodal investigations beyond simply a test of drug efficacy. While this hypothesis is indeed a central concern, the moxifloxacin case shows how there are numerous other methodological issues being tested si-

multaneously. When is a phase 2 trial justified? What explains a lack of efficacy in African patients? What is the predictive relationship between animal model results and human trials?

The particular nature of the discordance in phase 2, for example, also raises an array of questions about appropriate experimental design for tuberculosis trials. Despite the discordance over the primary result (i.e., moxifloxacin's effectiveness for shortening tuberculosis treatment time), four of the five studies still demonstrated at least one robust result: treatment with moxifloxacin is associated with increased culture conversions at time points earlier than 8 weeks. Yet, only two of the five studies used this as the primary endpoint. Rustomjee et al. (2008) justify this alternative surrogate with an appeal to two articles on tuberculosis-HIV/AIDS coinfection, whereas Burman et al. (2006) justify their selection of the 8-week culture conversion endpoint (as well as the power calculation in their study) with reference to earlier tuberculosis research on pyrazinamide, one of the drugs in the standard regimen. Pyrazinamide's addition to the regimen shortened treatment times by 3 months and increased 8-week conversion rates by an average of 13% (Burman et al. 2006, 332).

It is not obvious that either of these justifications is sufficient to set a methodological standard for tuberculosis research, but this is not to accuse either study (or both) of poor design. The use of alternative surrogates, as well as conflicts about which is preferable, merely speaks to a research program for which the underlying methodology is still in flux. In other words, the gold standard of phase 2 design for tuberculosis testing has not yet been determined. And this illuminates another critical function of robustness analysis: to identify these deeper methodological questions for further investigation. Multimodal discordance may be epistemically problematic for some hypotheses (e.g., is moxifloxacin promising enough to warrant phase 3 trials?) and yet epistemically rich for others (e.g., what is the more predictive surrogate endpoint in phase 2 tuberculosis trials?). Future success for moxifloxacin regimens against tuberculosis would suggest that time to conversion is the more predictive surrogate; future failure would suggest that culture negative at time t is the more predictive.

This shift in emphasis from thinking about robustness in terms of support for a single hypothesis to thinking about it in terms of a process of multihypothesis testing and methodological refinement further undercuts the worry that "robustness-style arguments do not tell us what to believe in situations of evidential discordance." The target belief that the multimodal evidence is taken to support is not necessarily a theoretical or causal belief, for example, the true effect of the drug or the truth of the posited causal relationship between drug administration and disease modification. There are beliefs about the relationships among testing modalities, the best designs for experimental modalities, the optimal strategies for a program of

multimodal testing, and so on. All of these different beliefs are relevant to interpreting the experimental outcomes of multimodal investigations, whether concordant or discordant.

We can think of this as the constructive lesson from the problem of underdetermination: disconfirmation and failures of multimodal robustness are hypothesis generating. It is therefore the pursuit of a robust result that is critical to scientific epistemology. The support for a hypothesis or the property of robustness is the desired end state, but intermediate stages (even of indeterminate length) where multimodal evidence is discordant and the degree of support (or even the fundamental methodology) is called into question do not undermine the epistemological value of robustness arguments. On the contrary, robustness is the philosophical concept—and robustness analysis, the methodological principle—that drives the research program forward.

6. Conclusion. The value of multimodal evidence is a cornerstone of scientific epistemology. Along with replication of results, using multiple means of testing a hypothesis is often taken as the primary method by which science is able to distinguish what is real from what is illusory. Yet, it is important to distinguish the philosophical implications of those "easy" examples, where multimodal evidence is concordant, from those messier examples where multimodal evidence is discordant.

Despite its flaws, Stegenga's argument is successful in drawing attention to some of the challenges that discordant evidence presents for the philosophical understanding of robustness. We should not assume that most instances of multimodal evidence are concordant, nor should we assume that it is trivial to amalgamate different evidential modalities toward a clear state of belief. Clinical research, in general, and the moxifloxacin case, in particular, vividly illustrate these points. Given that we demand a demonstration of multimodal robustness of medical treatments, it is important to clarify the philosophical understanding of this concept.

I have argued that the goal of identifying robust results and a commitment to robustness analysis are methodological mechanisms that drive research forward. Thus, Stegenga's criticism that robustness does not "tell us what to believe" is largely misplaced. The critical function of robustness in cases of discordance is not to tell researchers what to believe, but rather to generate new hypotheses. As we saw, although moxifloxacin's efficacy for treating tuberculosis remained uncertain, the discordant evidence generated an array of explanatory and methodological questions, many of which were later followed up and tested in subsequent investigations.

The lesson here underscores the importance of the distinction I offered in the introduction between robustness, understood as a property of hypotheses or results, and robustness analysis, understood as a systematic search

for those robust hypotheses and results. The problem of discordance points to complications in judging robust hypotheses, but it does not undermine the value of robustness analysis.

REFERENCES

Burman, William J., et al. 2006. "Moxifloxacin versus Ethambutol in the First 2 Months of Treatment for Pulmonary Tuberculosis." *American Journal of Respiratory and Critical Care Medicine* 174:331–38.

Chapin, Charles V. 1910. *The Sources and Modes of Infection*. New York: Wiley.

Conde, Marcus B., et al. 2009. "Moxifloxacin versus Ethambutol in the Initial Treatment of Tuberculosis: A Double-Blind, Randomised, Controlled Phase II Trial." *Lancet* 373:1183–89.

Culp, Sylvia. 1994. "Defending Robustness: The Bacterial Mesosome as a Test Case." In *PSA 1994: Proceedings of the 1994 Biennial Meeting of the Philosophy of Science Association*, ed. David Hull, Micky Forbes, and Richard M. Burian, 46–57. East Lansing, MI: Philosophy of Science Association.

Dorman, Susan E., et al. 2009. "Substitution of Moxifloxacin for Isoniazid during Intensive Phase Treatment of Pulmonary Tuberculosis." *American Journal of Respiratory and Critical Care Medicine* 180:273–80.

Franklin, Allan. 2002. *Selectivity and Discord: Two Problems of Experiment*. Pittsburgh: University of Pittsburgh Press.

Freedman, Benjamin. 1987. "Equipoise and the Ethics of Clinical Research." *New England Journal of Medicine* 317:141–45.

Gillespie, Stephen, and O. Billington. 1999. "Activity of Moxifloxacin against Mycobacteria." *Journal of Antimicrobial Chemotherapy* 44:393–95.

Gillespie, Stephen H., Roly D. Gosling, Leonard Uiso, Noel E. Sam, Esther G. Kanduma, and Timothy D. McHugh. 2005. "Early Bactericidal Activity of a Moxifloxacin and Isoniazid Combination in Smear-Positive Pulmonary Tuberculosis." *Journal of Antimicrobial Chemotherapy* 56:1169–71.

Gosling, Roly D., Leonard O. Uiso, Noel E. Sam, Emily Bongard, Esther G. Kanduma, Mramba Nyindo, Richard W. Morris, and Stephen H. Gillespie. 2003. "The Bactericidal Activity of Moxifloxacin in Patients with Pulmonary Tuberculosis." *American Journal of Respiratory and Critical Care Medicine* 168:1342–45.

Hall, Caroline B. 2007. "The Spread of Influenza and Other Respiratory Viruses: Complexities and Conjectures." *Clinical Infectious Diseases* 45:353–59.

Hay, Michael, David W. Thomas, John L. Craighead, Celia Economides, and Jesse Rosenthal. 2014. "Clinical Development Success Rates for Investigational Drugs." *Nature Biotechnology* 32:40–51.

Hey, Spencer P., Charles M. Heilig, and Charles Weijer. 2013. "Accumulating Evidence and Research Organization (AERO) Model: A New Tool for Representing, Analyzing, and Planning a Translational Research Program." *Trials* 14:159.

Ji, Baohong, Nacer Lounis, Caroline Maslo, Chantal Truffot-Pernot, Pascale Bonnafous, and Jacques Grosset. 1998. "In Vitro and In Vivo Activities of Moxifloxacin and Clinafloxacin against Mycobacterium Tuberculosis." *Antimicrobial Agents and Chemotherapy* 42:2066–69.

Johnson, J., et al. 2006. "Early and Extended Early Bactericidal Activity of Levofloxacin, Gatifloxacin and Moxifloxacin in Pulmonary Tuberculosis." *International Journal of Tuberculosis and Lung Disease* 10:605–12.

Levins, Richard. 1966. "The Strategy of Model Building in Population Biology." *American Scientist* 54:421–31.

———. 1993. "A Response to Orzack and Sober: Formal Analysis and the Fluidity of Science." *Quarterly Review of Biology* 68:547–55.

Lounis, Nacer, Abdelhalim Bentoucha, Chantal Truffot-Pernot, Baohong Ji, Richard J. O'Brien, Andrew Vernon, Giorgio Roscigno, and Jacques Grosset. 2001. "Effectiveness of Once-

Weekly Rifapentine and Moxifloxacin Regimens against Mycobacterium Tuberculosis in Mice." *Antimicrobial Agents and Chemotherapy* 45:3482–86.

Mac Kenzie, William R., et al. 2011. "Geographic Differences in Time to Culture Conversion in Liquid Media: Tuberculosis Trials Consortium Study 28; Culture Conversion Is Delayed in Africa." *PLoS One* 6:e18358.

Mateo, Joaquin, Michael Ong, David S. P. Tan, Michael A. Gonzalez, and Johann S. de Bono. 2013. "Appraising Iniparib, the Parp Inhibitor That Never Was—What Must We Learn?" *Nature Reviews Clinical Oncology* 10:688–96.

Miyazaki, Eishi, Miki Miyazaki, Jong Min Chen, Richard E. Chaisson, and William R. Bishai. 1999. "Moxifloxacin (bay12-8039), a New 8-Methoxyquinolone, Is Active in a Mouse Model of Tuberculosis." *Antimicrobrial Agents and Chemotherapy* 43:85–89.

Nuermberger, Eric L., Tetsuyuki Yoshimatsu, Sandeep Tyagi, Richard J. O'Brien, Andrew N. Vernon, Richard E. Chaisson, William R. Bishai, and Jacques H. Grosset. 2004. "Moxifloxacin-Containing Regimen Greatly Reduces Time to Culture Conversion in Murine Tuberculosis." *American Journal of Respiratory and Critical Care Medicine* 169:421–26.

Orzack, Steven H., and Elliott Sober. 1993. "A Critical Assessment of Levins's 'The Strategy of Model Building in Population Biology' (1966)." *Quarterly Review of Biology* 68:533–46.

Pletz, Mathias W. R., Andres De Roux, Andreas Roth, Karl-Heinz Neumann, Harald Mauch, and Hartmut Lode. 2004. "Early Bactericidal Activity of Moxifloxacin in Treatment of Pulmonary Tuberculosis: A Prospective, Randomized Study." *Antimicrobial Agents and Chemotherapy* 48:780–82.

Rasmussen, Nicolas. 1993. "Facts, Artifacts, and Mesosomes: Practicing Epistemology with the Electron Microscope." *Studies in History and Philosophy of Science* A 24:227–65.

Rustomjee, R., et al. 2008. "A Phase II Study of the Sterilising Activities of Ofloxacin, Gatifloxacin and Moxifloxacin in Pulmonary Tuberculosis." *International Journal of Tuberculosis and Lung Disease* 12:128–38.

Shandil, Radha K., Ramesh Jayaram, Parvinder Kaur, Sheshagiri Gaonkar, B. L. Suresh, B. N. Mahesh, R. Jayashree, Vrinda Nandi, Sowmya Bharath, and V. Balasubramanian. 2007. "Moxifloxacin, Ofloxacin, Sparfloxacin, and Ciprofloxacin against Mycobacterium Tuberculosis: Evaluation of In Vitro and Pharmacodynamic Indices That Best Predict In Vivo Efficacy." *Antimicrobial Agents and Chemotherapy* 51:576–82.

Soler, Léna, Emiliano Trizio, Thomas Nickles, and William Wimsatt, eds. 2012. *Characterizing the Robustness of Science: After the Practice Turn in Philosophy of Science*. Dordrecht: Springer.

Staley, Kent W. 2004. "Robust Evidence and Secure Evidence Claims." *Philosophy of Science* 71:467–88.

Stass, H., A. Dalhoff, D. Kubitza, and U. Schühly. 1998. "Pharmacokinetics, Safety, and Tolerability of Ascending Single Doses of Moxifloxacin, a New 8-Methoxy Quinolone, Administered to Healthy Subjects." *Antimicrobial Agents and Chemotherapy* 42:2060–65.

Steenwinkel, Jurriaan E. M., Gerjo J. de Knegt, T. Marian, Alex van Belkum, Henri A. Verbrugh, Kristin Kremer, Dick van Soolingen, and Irma A. J. M. Bakker-Woudenberg. 2010. "Time-Kill Kinetics of Anti-tuberculosis Drugs, and Emergence of Resistance, in Relation to Metabolic Activity of Mycobacterium Tuberculosis." *Journal of Antimicrobial Chemotherapy* 65:2582–89.

Stegenga, Jacob. 2009. "Robustness, Discordance, and Relevance." *Philosophy of Science* 76:650–61.

———. 2012. "Rerum Concordia Discors: Robustness and Discordant Multimodal Evidence." In Soler et al. 2012, 207–26.

Wang, J. Y., J. T. Wang, T. Tsai, C. Hsu, C. Yu, P. Hsueh, L. Lee, and P. Yang. 2009. "Adding Moxifloxacin Is Associated with a Shorter Time to Culture Conversion in Pulmonary Tuberculosis." *International Journal of Tuberculosis and Lung Disease* 14:65–71.

Weber, Thomas P., and Nikolaos I. Stilianakis. 2008. "Inactivation of Influenza A Viruses in the Environment and Modes of Transmission: A Critical Review." *Journal of Infection* 57:361–73.

Weijer, Charles, and Paul B. Miller. 2004. "When Are Research Risks Reasonable in Relation to Anticipated Benefits?" *Nature Medicine* 10:570–73.

Wimsatt, William. 1980. "Reductionist Research Strategies and Their Biases in the Units of Selection Controversy." In *Scientific Discovery: Case Studies*, ed. Thomas Nickles, 213–59. Dordrecht: Reidel.

———. 1981. "Robustness, Reliability and Overdetermination." In *Scientific Inquiry and the Social Sciences*, ed. M. Brewer, 124–63. San Francisco: Jossey-Bass.

———. 2007. *Re-engineering Philosophy for Limited Beings: Piecewise Approximations to Reality.* Cambridge, MA: Harvard University Press.

Yoshimatsu, Tetsuyuki, Eric Nuermberger, Sandeep Tyagi, Richard Chaisson, William Bishai, and Jacques Grosset. 2002. "Bactericidal Activity of Increasing Daily and Weekly Doses of Moxifloxacin in Murine Tuberculosis." *Antimicrobial Agents and Chemotherapy* 46:1875–79.