Theory Testing and Implication in Clinical Trials

March 1, 2014

Abstract

John Worrall (2010) and Nancy Cartwright (2011) argue that randomized controlled trials (RCTs) are "testing the wrong theory." RCTs are designed to test inferences about the causal relationships in the study population, but this does not guarantee a justified inference about the causal relationships in the more diverse population in clinical practice. In this essay, I argue that the epistemology of theory testing in trials is more complicated than either Worrall's or Cartwright's accounts suggest. I illustrate this more complex theoretical structure with case-studies in medical theory testing from (1) Alzheimer's research and (2) anti-cancer drugs in personalized medicine.

1 Introduction

John Worrall (2010) and Nancy Cartwright (2011) have both argued that there is a mismatch between the theory being tested in a randomized controlled clinical trial (RCT) and the theory that medical practitioners are actually interested in. Worrall describes this as the problem of external validity: An RCT may support internally valid inferences about the causal relationships in the study population, but this does not guarantee a justified inference about the causal relationships in the target population of interest—i.e., the usually more diverse patient population that physicians actually encounter in the clinic. Since it is the causal relationship between treatment and patient outcome in this more diverse population that we ultimately care about, the RCT is "testing the wrong theory" (p.361). Or as Cartwright puts it: "an RCT supports only an 'it-works-somewhere' claim," but what we need are justified "it will work for us" claims (p.1401).

There is something fundamentally correct in both Cartwright's and Worrall's arguments: Many of the experimental components used in RCTs are designed to secure internal validity at the expense of external validity. Yet, the epistemological relationship between translational clinical trials (whether randomized or not) and the underlying scientific theories is more complicated and, in some ways, more subtle than either of their accounts would suggest. In this essay, I will illustrate this more complex theoretical structure through two examples, drawn from trials in Alzheimer's research and anti-cancer drugs for personalized medicine. I argue that the more complicated epistemology vis-a-vis theory testing revealed in these cases illuminates how Worrall's and Cartwright's philosophical conclusion relies on an overly narrow conception of what trials can show.

2 Testing the Wrong Theory

Let us begin, as Worrall does, by imagining an RCT that is evaluating some new intervention S as a treatment for some condition C in a sample population P. Supposing that the trial is positive, what can we now conclude? Worrall cautions that we should not conclude the "dangerously vague" claim that "S is effective for treating C". Rather, the trial's result only warrants a narrower claim that "S when administered in a very particular way to a very particular set of patients for a particular length of time is more effective than some comparator treatment" (Worrall, 2010, p.361).

In other words, if we assume that the trial is internally valid, then it only justifies the claim that "S was effective for condition C in population P". But the practicing physician needs to know how or whether this effectiveness claim generalizes to the variations on these parameters that she is likely to encounter in the clinic. Does the relationship between S and C also hold for more elderly patients (who are typically excluded from trials)? Does it hold if we modify the dose and schedule to accommodate patients with co-morbidities or concomitant medications? Is S also an effective intervention for treating the related condition, C^*? All of these questions speak to the problem of external validity: An RCT may demonstrate a causal relationship between S and C in the studied population P. But what we want to know is whether this relationship holds for $\mathbb{S}$ and $\mathbb{C}$ in the population $\mathbb{P}$—where $\mathbb{S}$, $\mathbb{C}$, and $\mathbb{P}$ are the respective sets of plausible variations on the intervention, condition, and patient populations that the physician encounters in clinical practice.

And as Worrall goes on to point out, there are often reasons to doubt that P is a good representative (or perhaps even a member) of the relevant clinical set $\mathbb{P}$. For example, the limited observation time in an RCT weakens inferences about the safety of treatments for chronic diseases, such as diabetes or arthritis. Even though large RCTs for these conditions will usually include a few years of follow-up, the general patient population is likely to be taking these medications for decades, and the RCT has not provided any evidence about such long-term effects. Similarly, some RCTs will include time-consuming procedures or expensive diagnostics as a part of the treatment regimen. Insofar as these same procedures or tests are unavailable to the clinician (due to excessive cost, timing, or feasibility), then the results of the RCT can fail to provide clinically relevant evidence about how the intervention should be used in practice.

Cartwright (2011) extends this line of argument with some additional analytic resources. She draws a distinction between experimental "vouchers" and ”clinchers" (cf. Cartwright, 2007). A "voucher" is an experiment that renders its conclusion more probable, whereas a "clincher" is an experiment that deductively implies its conclusion. As she defines it, an ideal RCT "where all the requisite premises are met" is a clincher—and this is presumably why the RCT sits atop the hierarchy of evidence-based medicine.

But what are these "requisite premises" on which the "clinching" depends? Cartwright enumerates three of them: (1) probabilistic dependence calls for causal explanation; (2) all causal features in the study population P relevant to the outcome, except for the treatment, are equally distributed between the treatment and control arms; and (3) the experimental treatment S is the only plausible explanation for the observed difference in outcome between the arms (p.1400).[1]

Let us set premise (1) aside here, since discussing the necessity (or not) of causal explanations for probabilistic dependence will take us too far afield. Premises (2) and (3) deserve some attention, however. As Cartwright acknowledges, RCTs are explicitly designed to satisfy these two claims. Random treatment allocation, in particular, is typically justified for exactly this reason: It controls for all known and unknown confounders in the study population. Restrictive eligibility criteria, strict treatment protocols, allocation concealment, and blinded outcome assessment are also characteristic features of the RCT—all of which are introduced to eliminate the influence of bias, and thereby increase our confidence in premises 2 and 3.

But as Cartwright observes, these methodological steps also render the RCT less like clinical practice. In the clinic, physicians will often modify a treatment's dose or schedule, or even switch patients from one drug to another, in the face of observed inefficacy, adverse reactions, or side-effects. Patients are also neither blinded to their prescribed treatment nor prescribed a treatment randomly. And just as Worrall argued, the clinical patient population $\mathbb{P}$ is usually far more diverse than the study population P. Each of these differences between the RCT and practice weaken the inference (i.e., generalizability) from causal claims about what occurred in the study to causal claims about what will occur in the clinic.

To resolve this problem, Cartwright argues that we need justified claims about the causal "capacities" of our treatment S—that is, *theoretical* warrant for thinking that S is a good representative of $\mathbb{S}$, C a good representative of $\mathbb{C}$, and P a good representative of $\mathbb{P}$. As she puts it, we need the theoretical understanding of "why the treatment should have the power to produce the outcome". Unfortunately, all we get from an RCT is evidence that the S can "work somewhere," but this is not the same as having a justified theory of why we should expect that "it will work for us" (p.1401).

3 External Validity and Underdetermination

As I suggested above, there is something fundamentally correct in Worrall's and Cartwright's arguments. The RCT is typically designed to ensure internal validity—to "clinch" (to use Cartwright's term) the causal hypothesis that the experimental treatment S is efficacious against condition C in the study's population P. But as we just saw, the steps taken to guarantee the validity of this inference will often weaken its external validity. And this trade-off

[1] Cartwright does not use these variables in her formulation. I introduce them here to better accord with the Worrall discussion above, but I trust it does no violence to her account. I have, however, weakened her third premises. Her original wording—"the *only explanation possible* is that the treatment caused the outcome" (p.1400, emphasis added)—is far too strong and inconsistent with any credible account of RCT methodology, see for example Shadish et al. (2002) or Friedman et al. (2010).

on internal versus external validity leads us to Worrall and Cartwright conclusions: RCTs are not testing the right theory. They are not telling us what we need to know.

This conclusion is consonant with others in the medical literature who have called for more "comparative effectiveness" trials. For example, Tunis et al. (2003) argue that too many RCTs evaluate the new drug against a placebo comparator, even when there is already a proven effective treatment available. But if the only evidence about some new drug, A, is its superiority to placebo, this does not provide clinicians with sufficient knowledge about whether they should be prescribing A over the old standard of care. It can also be traced back to Schwartz and Lellouch (1967), who drew a distinction between *pragmatic* and *explanatory* trials. A pragmatic trial is conducted under conditions similar to clinical practice and seeks to answer a question about medical decision-making, e.g., "Which treatment should physicians use in practice?" Whereas an explanatory trial is conducted under "ideal" scientific conditions and seeks to answer a question about scientific understanding, e.g., "What is the true biological effect of drug S?" Their philosophical point is similar to Worrall's and Cartwright's: Different experiments can address different theories, so we should be conducting experiments that answer the questions we are actually interested in. If we want to answer clinically relevant questions, then we should be conducting pragmatic trials that maximize external validity.

I take it, however, that Worrall and Cartwright are not simply echoing Schwartz and Lellouch. They seem to be saying something stronger—namely, that RCTs are testing the "wrong theory." But what should we make of this claim? It is certainly true that most RCTs adopt some version of Neyman-Pearson hypothesis testing, and are therefore, strictly speaking, tests of a single hypothesis (or single theory, if you prefer). Yet, it would be a mistake to think that an RCT has no further theoretical importance. This much follows trivially from underdetermination: Multiple scientific theories are involved in the design of an experiment and therefore multiple scientific theories are implicated by the evidence produced—e.g., theories about the therapeutic class (of which the drug is just one member) and its relationship to disease modification; theories about the diagnostic assays and their relationship to the disease prognosis; theories of disease ontology and pathophysiology. A negative RCT, for example, does not necessitate that the researchers reject any causal link between the treatment S and the condition C. Perhaps S will be effective against C in a slightly different population P^*. The essential point of underdetermination is that there are always auxiliary hypotheses or other theoretical modifications that can be made to accommodate the evidence.[2]

Since it seems unlikely that Worrall and Cartwright would object to the relevance of theoretical underdetermination in RCTs, perhaps we ought to interpret their conclusion differently. Maybe what they are really arguing is that the inferences to these other theories are not well-justified by the evidence produced in an RCT. Cartwright, in particular, has the conceptual resources to still conclude that RCTs are not "clinchers". At best, they are only "vouchers" for most of the relevant theoretical claims.

[2] Worrall discusses Duhem's problem earlier in the same article (Worrall, 2010, p.358), but overlooks its relevance for his argument on theory testing. See also Anderson (2006), Howick (2009), and Hey and Weijer (2013) for more details discussions of Duhem's problem and its importance for understanding the methodology of clinical trials.

But if this is really their conclusion, much of the philosophical force of their argument is lost. If the argument simply is the same as Schwarz and Lellouch's or Tunis et al.'s—that we need pragmatic trials to answer (or "vouch for") clinically relevant questions—then I agree entirely. But this does not seem consistent with much of what Worrall and Cartwright argue. Even the most pragmatic trial is still just a "voucher" for clinical effectiveness. It is asking the more clinically relevant question. Yet, if it differs in any way from the clinical setting, then on Worrall's and Cartwright's view, it does not provide the right kind of evidential support.

Indeed, Cartwright emphasizes that "RCTs do not, without a series of strong assumptions, warrant predictions about what happens in practice" (p.1400). And Worrall concludes that we should really be doing observational studies rather than RCTs (p.362). But these conclusions seem untenable. Even setting aside the obvious objection to Worrall that observational studies have their own biases and methodological limitations (and are just as subject to underdetermination), it is much to strong to demand of an experiment that it provide direct evidence or causal certainty before we can draw externally valid inferences from it. "Clinchers" may be a worthy philosophical ideal, but it does not follow that this is a plausible experimental benchmark. Demanding deductive causal certainty is to ask more of an experiment than it can plausibly provide.

4 Trials and Theoretical Implication

So where does this leave us? I agree with Worrall and Cartwright that the external validity of RCTs is limited (as it is for every experiment). Yet, the story of clinical trials and theory testing is more complicated than either of their accounts would seem to suggest. In this section, I will discuss two examples of theory testing in medical research, each of which illuminates a number of different ways in which trials have a theoretical import beyond their specific testing hypothesis.

4.1 The Amyloid Cascade

Much of Alzheimer's research has been driven by a mechanistic theory of the amyloid cascade, which posits that the characteristic neurodegeneration of Alzheimer's disease is caused by amyloid-β plaque accumulation in the brain. However, as Karran et al. (2011) describe, even as various "amyloid-centric" approaches have failed (e.g., the drugs tramiprosate, tarenflurbil, semagacestat all failed in development), the fundamental amyloid cascade theory has not been rejected. It has only been modified. They now distinguish between three different theoretical amyloid-centric strategies: reducing amyloid-β production, facilitating amyloid-β clearance, and preventing amyloid-β aggregation (p.700).

The drug company, Genentech's, anti-amyloid monoclonal antibody, crenezumab, is one such amyloid-centric drug currently undergoing clinical trials. In fact, it is being tested in two different Alzheimer's trials: One is a long-term single arm trial in patients with mild to moderate Alzheimer's symptoms; the other is a double-blind RCT testing crenezumab as a neuroprotective agent in a genetically homogeneous population in Columbia.[3] So what are the theoretical implications of these trials?

For the single arm study, a negative result would provide evidence that crenezumab is not an effective strategy for treating Alzheimer's symptoms. It would also provide evidence that similar monoclonal antibodies are unlikely to be effective, as well as further evidence for the growing suspicion that once amyloid-β deposition has begun, removing amyloid-β is unlikely to offer any therapeutic benefit (Golde et al., 2011). Whereas a positive result would confirm both (a) that crenezumab and similar monoclonal antibodies may be viable strategies, and (b) that an amyloid-β clearance strategy is effective.

Similarly, for the RCT in Columbia, a negative result would be evidence against crenezumab's effectiveness. It would also provide disconfirming evidence that preventing amyloid-β aggregation offers any neurodegenerative protection. A positive result would confirm both of those theories: preventing amyloid-β aggregation is a viable strategy and crenezumab, in particular, is likely to be an effective treatment.

I am happy to grant a possible Cartwright objection here that neither of these trials "clinch" any of these theoretical claims. But surely the more relevant question is whether or not these trials provide sufficient evidence for informing clinical decision-making. And on that point it is instructive to observe that part of the inclusion criteria for the Columbian RCT is that all patient-subjects must be carriers of a specific gene mutation (PSEN1 E280A), which is known to cause early-onset Alzheimer's disease (cf. Belluck, 2012). Supposing that this trial has a positive result, what are clinicians justified in concluding about other patient populations at risk for Alzheimer's? Worrall's and Cartwright's arguments imply that clinicians would still lack sufficient evidence for prescribing crenezumab outside of that specific genetic population. But the theoretical warrant from these trials is not so weak. A success for crenezumab lends evidential support for a range of theoretical propositions, some of which would be sufficient to justify a clinician's decision to prescribe an approved anti-amyloid agent for her Alzheimer's patients.

And it brings us to the heart of the issue: The directly tested theory in the Columbian RCT could be thought of as resembling the narrow proposition, much as Worrall originally construed it: "Crenezumab (S) is effective for preventing the development of Alzheimer's disease (C) in the Columbian patient population possessing the PSEN1 E280A genetic mutation (P)." But this does not exhaust the theoretical relevance of the trial. Whatever its final result—but particularly if it is positive—researchers and clinicians will be in a better position to draw valid (albeit inductive) inferences about future preventative strategies against Alzheimer's. Specifically, they would be justified in inferring potential efficacy for other preventative anti-amyloid interventions ($\mathbb{S}$); extrapolating the strategy for related conditions, such as sporadic Alzheimer's ($\mathbb{C}$); or prescribing anti-amyloid medications for other patient populations at high-risk for developing amyloid-related neurodegenerative diseases ($\mathbb{P}$). To be sure, these would all still be inferences with some degree of causal uncertainty, but it does not follow from the lack of certainty that the inferences are unwarranted or unjustified. On the contrary, if an anti-amyloid strategy is shown to be effective in an RCT, it would arguably

[3] See http://clinicaltrials.gov/ct2/results?term=Crenezumab&Search=Search, retrieved February 27, 2014.

violate the physicians' duty of care to withhold the treatment.

4.2 Personalized Cancer Medicine

In many ways, the amyloid cascade and Alzheimer's case is an exemplar for the traditional model of clinical translation, where the driving theories concern the experimental drug's effectiveness and the mechanism of disease. As we saw, a new Alzheimer's drug that is successfully vetted in trials is taken to confirm the particular drug's effectiveness, the effectiveness of the strategic class, and the underlying theories of disease pathophysiology. Whereas the drug's failure can be attributed to either a problem with one of these theories, a faulty auxiliary hypothesis, or an operational error in one or more of the experiments.

The development of new personalized medicines (PM), however, is not well-characterized by this model. The goal in PM is to equip the health-care system with an array of clinically validated diagnostics, each of which would allow physicians to test their patients for the presence or absence of a particular biomarker (e.g., a genetic mutation in their tumor specimen), and then use these results to tailor decision-making about the appropriate course of treatment. If successfully implemented, these biomarker diagnostics would potentially save the health-care system billions of dollars and prevent needless patient suffering due to futile interventions.

On its face, the epistemology of PM, in some ways, better accommodates Worrall's and Cartwright's views. That is, PMs are designed to be effective in a very narrowly defined patient population—i.e., only those patients with the specific biomarker. Thus, the study population in RCTs for PM is far more likely to resemble the target population in clinical practice. However, in contrast to the traditional model of medical research and drug development, which hinges on effective therapeutic agents, the promise of PM largely depends upon the development of high-quality biomarker diagnostics. And this further complicates the theoretical implications of PM trials.

Consider the case of the alkylating agent, temozolomide. This drug was derived from the older, widely-used (although quite toxic) cancer agent, dacarbazine, and works by attaching an alkyl group to the cancer cell DNA, disrupting its growth and leading to cell death. Interestingly, despite sharing the same mechanism of action, temozolomide and dacarbazine are used in differenct cancers. Dacarbazine is approved for use against Hodgkin lymphoma and melanoma; temozolomide is approved for use against anaplastic astrocytoma and glioblastoma multiforme.

Let us label this broadly defined mechanistic theory of using alkylating cancer drugs, T_1:

T_1 Alkylating agents ($S_1 \ldots S_n$) are a viable treatment strategy for some patient populations ($P_1 \ldots P_n$) with some cancers ($C_1 \ldots C_n$).

Thus, dacarbazine and temozolomide are two of the agents in the set $S_1 \ldots S_n$, and the various cancers for which they have been approved are the members of the set $C_1 \ldots C_n$. One of the challenges in cancer treatment is that the patient population that benefits from a particular agent is not fully determined by their cancer-type. For example, not all patients with glioblastoma will benefit from temozolmide therapy. And this is where diagnostic biomarker assays come into play. Indeed, the theory underlying all of PM is that there are genetic markers in a patient's tumor which can predict whether or not they are likely to benefit from a treatment.

One of the proposed biomarkers for temozolomide is the DNA repair gene O-6-methylguanine-DNA methyltransferase, typically abbreviated as "MGMT". A landmark study by Hegi et al. (2004) identified a positive correlation between patient tumor response to temozolomide therapy and high levels of methylated MGMT expression in their tumor specimens. Their conclusion can be characterized by the more specific theoretical hypothesis, T_2:

T_2 Temozolomide chemotherapy (S_g) is most likely to be effective against glioblastoma tumors (C_g) for those patients whose tumor specimens express high levels of methylated MGMT (P_g).

We can think of T_2 as a sub-theory of T_1, since it describes a relationship among a single triad of the treatment-condition-population parameters. And although T_1 is uncontroversial and has already been taken up in clinical practice, T_2 is still being rigorously evaluated in trials.[4] But just as we saw with the amyloid cascade theory and crenezumab, a positive or negative result in any of these trials has theoretical implications for both T_1 and T_2.

Yet, many of these trials have an additional dimension of uncertainty derived from the predictivity of the diagnostic assay (or assays) used to assess the methylated MGMT biomarker. There are multiple techniques that can be used to determine the level of methylated MGMT in a specimen and these different techniques do not all discriminate the glioblastoma patients in the same way. In effect, they each define the target population P_g differently. One recent study, for example, compared the sensitivity and specificity of a methylation-specific polymerase chain reaction (MS-PCR) assay, which amplifies the relevant CpG islands of the tumor specimen's DNA, against an immunohistochemistry staining (IHC) assay, which assesses the reactivity of tumor cells against a specific antibody (Lechapt-Zalcman et al., 2012). They found that although both assays positively predicted benefit from temozolomide therapy, the agreement between them was only about 70%. That is, 30% of the samples tested positive for methylated MGMT on one test, but negative on the other.

This makes the recommendation for clinical practice more problematic. What is the true patient population for our theory T_2? Is it the patients whose samples test positive on MS-PCR or IHC? A clinician's decision to recommend temozolomide now hinges, in part, on their selection of assay.

Lechapt-Zalcman et al. (2012) attribute this discrepancy largely to false-positives with the IHC, which on its face, would seem to suggest that MS-PCR is the better assay for defining the population P_g (p.4553). But they also note that the accuracy of MS-PCR depends upon high-quality cryopreserved tumor specimens, which is expensive and not widely available in the clinical setting (p.4552). Thus, despite its being the less accurate of the two assays, an IHC assay may be the more clinically useful diagnostic. And this brings us back to the problem of external validity. If MS-PCR is too expensive and unlikely to be used in the clinic, then the Worrall or Cartwright arguments would suggest that future trials ought to only

[4] At the time of this writing, 10 trials are registered on clinicaltrials.gov examining the implications of temozolomide and MGMT for the treatment of glioblastoma.

investigate IHC. Since IHC is the technique available to clinicians, then presumably, what clinicians want is evidence about its capacity to delineate the responding patient population P_g.

Unfortunately, even this seemingly reasonable suggestion still relies on an oversimplification of the theoretical implications in these studies. To wit, we should observe that the effective use of a diagnostic test depends on knowing its misclassification rate, i.e., the false-positive and false-negative error rates. If a gold-standard diagnostic exists—that is, a diagnostic with perfect sensitivity and specificity—then these error rates are easy to determine. One can simply compare the classification of the imperfect diagnostic, e.g., IHC, to the classification according to the gold-standard. Of course, in practice, there are no gold-standards. Every diagnostic is imperfect. However, there are validated techniques for accurately estimating the error rates of a test on the basis of multiple diagnostics. In essence, these are robustness strategies, which use multiple independent (or sometimes conditionally dependent) tests in order to arrive at estimates for the error rates of each individual diagnostic (Joseph et al., 1995).

And indeed, relying on multiple diagnostics is precisely the strategy adopted in some of the more recent temozolomide studies (cf. Lalezari et al., 2013). Given that IHC is known to be inaccurate, researchers in these trials can use other, more accurate diagnostics (e.g., MS-PCR, pyrosequencing) in combination with IHC in order to derive better estimates for the error rates when using a single IHC diagnostic test. These estimates can then be used by clinicians, who may only have access to one diagnostic method, to make informed decisions about their patient's true biomarker status and potential benefit from temozolomide therapy.

The essential philosophical point here is that these rigorous biomarker studies do have weaker external validity. They employ multiple diagnostics and robustness strategies, which may be unavailable or unwieldy in clinical practice. Yet, their use of multiple diagnostics toward a more robust theoretical understanding of the various individual techniques is precisely what makes them informative for clinical practice. Contrary to Cartwright's claim, these explanatory (or "ideal") trials are addressing the clinically relevant theoretical question—"What is the accuracy of IHC for predicting response to temozolomide?" This is exactly the kind of information that clinicians need to know in order to make the most of PM in cancer.

5 Conclusion

What theory or theories are tested in clinical trials? I have argued here that the answer to this question is more complicated than suggested by either Worrall (2010) or Cartwright (2011) in their critiques of RCTs. Their emphasis on the problem of external validity is helpful, insofar as it draws greater attention to the need for studies that address clinically relevant questions. But their stronger conclusion against the theoretical warrant provided by RCTs relies on a significant oversimplification of trial epistemology.

As the problem of underdetermination entails, there are many theoretical implications of trials. The focal testing hypothesis of the form "Treatment S is effective for condition C in population P" is but one of the many theoretical claims that can be justifiably confimred, modified, or refuted in light of an trial's result. RCTs also generate evidence that is relevant for general theories about the viability of the mechanistic strategy, or the underlying pathophysiological theories of the disease, or the theories concerning biomarkers, diagnostic assays, and the predictive relationship that these bear to patient prognosis and treatment. All of these moving theoretical parts are potentially implicated. To suggest otherwise assumes an overly narrow and untenable view about what RCTs can show.

References

Anderson, J. A. (2006). The ethics and science of placebo-controlled trials: Assay sensitivity and the duhem-quine thesis. *Journal of Medicine and Philosophy 31*, 65–81.

Belluck, P. (2012). New drug trial seeks to stop alzheimers before it starts. *The New York Times*. May 15.

Cartwright, N. (2007). Are rcts the gold standard? *BioSocieties 2*(1), 11–20.

Cartwright, N. (2011). The art of medicine: A philosopher's view of the long road from rcts to effectiveness. *The Lancet 377*, 1400–1401.

Friedman, L. M., C. Furberg, and D. L. DeMets (2010). *Fundamentals of clinical trials* (4 ed.). Springer.

Golde, T. E., L. S. Schneider, and E. H. Koo (2011). Anti-aβ therapeutics in alzheimer's disease: the need for a paradigm shift. *Neuron 69*(2), 203–213.

Hegi, M. E., A.-C. Diserens, S. Godard, P.-Y. Dietrich, L. Regli, S. Ostermann, P. Otten, G. Van Melle, N. de Tribolet, and R. Stupp (2004). Clinical trial substantiates the predictive value of o-6-methylguanine-dna methyltransferase promoter methylation in glioblastoma patients treated with temozolomide. *Clinical Cancer Research 10*(6), 1871–1874.

Hey, S. P. and C. Weijer (2013). Assay sensitivity and the epistemic contexts of clinical trials. *Perspectives in Biology and Medicine 56*(1), 1–17.

Howick, J. (2009). Questioning the methodologic superiority of 'placebo' over 'active' controlled trials. *The American Journal of Bioethics 9*, 34–48.

Joseph, L., T. W. Gyorkos, and L. Coupal (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology 141*(3), 263–272.

Lalezari, S., A. P. Chou, A. Tran, O. E. Solis, N. Khanlou, W. Chen, S. Li, J. A. Carrillo, R. Chowdhury, J. Selfridge, et al. (2013). Combined analysis of o6-methylguanine-dna methyltransferase protein expression and promoter methylation provides optimized prognostication of glioblastoma outcome. *Neuro-oncology 15*(3), 370–381.

Lechapt-Zalcman, E., G. Levallet, A. E. Dugué, A. Vital, M.-D. Diebold, P. Menei, P. Colin, P. Peruzzy, E. Emery, M. Bernaudin, et al. (2012). O6-methylguanine-dna methyltransferase (mgmt) promoter methylation and low mgmt-encoded protein expression as prognostic markers in glioblastoma patients treated with biodegradable carmustine wafer implants after initial surgery followed by radiotherapy with concomitant and adjuvant temozolomide. *Cancer 118*(18), 4545–4554.

Schwartz, D. and J. Lellouch (1967). Explanatory and pragmatic attitudes in therapeutical trials. *Journal of chronic diseases 20*(8), 637–648.

Shadish, W. R., T. D. Cook, and D. T. Campbell (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage learning.

Tunis, S. R., D. B. Stryer, and C. M. Clancy (2003). Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. *Journal of the American Medical Association 290*(12), 1624–1632.

Worrall, J. (2010). Evidence: Philosophy of science meets medicine. *Journal of Evaluation in Clinical Practice 16*, 356–362.