Judging Quality and Coordination in Biomarker Diagnostic Development[*]

Spencer Phillips Hey

Received: 27/08/2014
Final Version: 31/03/2015

BIBLID 0495-4548(2015)30:2p.207-227
DOI: 10.1387/theoria.12697

ABSTRACT: What makes a high-quality biomarker experiment? The success of personalized medicine hinges on the answer to this question. In this paper, I argue that judgment about the quality of biomarker experiments is mediated by the problem of theoretical underdetermination. That is, the network of biological and pathophysiological theories motivating a biomarker experiment is sufficiently complicated that it often frustrates valid interpretation of the experimental results. Drawing on a case-study in biomarker diagnostic development from neurooncology, I argue that this problem of underdetermination can be overcome with greater coordination across the biomarker research trajectory. I then sketch an account for how coordination across a research trajectory can be evaluated. I ultimate conclude that what makes a high-quality biomarker experiment must be judged by the epistemic contribution it makes to this coordinated research effort.

Keywords: biomarkers, personalized medicine, underdetermination, neurooncology.

RESUMEN: ¿Qué convierte a un experimento con biomarcadores en un experimento de gran calidad? El éxito de la medicina personalizada depende de la respuesta a esta pregunta. En este artículo sostengo que el juicio sobre la calidad de los experimentos con biomarcadores está mediado por el problema de la subdeterminación teórica, es decir, la red de teorías biológicas y patofisiológicas que motivan un experimento con biomarcadores es lo bastante complicada como para frustrar a menudo una interpretación válida de los resultados experimentales. A partir de un caso de desarrollo de diagnóstico con biomarcadores, defiendo que el problema de la subdeterminación puede ser superado con mayor coordinación en la trayectoria de investigación sobre el biomarcador. Después sugiero un enfoque para evaluar la coordinación a lo largo de una trayectoria de investigación. Por último concluyo que lo que hace que un experimento con biomarcadores tenga una alta calidad debe dirimirse en función de la contribución epistémica que aquél realiza sobre este esfuerzo investigador coordinado.

Palabras clave: biomarcadores, medicina individualizada, subdeterminación, neurooncología.

1. Introduction

What makes a high-quality biomarker experiment? The success of personalized medicine hinges on the answer to this question. Since the ultimate goal of personalized medicine is to base treatment decisions on a patient's biomarker status (e.g., the presence or absence of a genetic mutation), success requires that physicians have access to reliable

[*] This research received support under the PACEOMICS project, funded by Genome Canada, Genome Quebec, Genome Alberta, and the Canadian Institutes for Health Research (CIHR). I would like to thank Georgina Freeman, Jonathan Kimmelman, Brianna Barsanti-Innes, the audience at the PSX4 Conference at the University of Pittsburgh, and the anonymous reviewers at this journal for helpful feedback on earlier versions of these ideas and this manuscript.

biomarker diagnostics—i.e., biological assays that can accurately detect the biomarker and thereby predict whether or not a patient is likely to benefit from a course of treatment. Indeed, a health-care system equipped with an array of valid biomarker diagnostics could prevent much unnecessary patient suffering and health-care spending due to futile interventions.

Unfortunately, the quality of most biomarker experiments to date has been quite low (Kyzas et al. 2005, McShane et al. 2005, Mallett et al. 2009). This has lead to a "vicious cycle," where the evidence from biomarker experiments is poorly valued, biomarker research is poorly funded, and the costs of diagnostics are not reimbursed by health-care providers. Consequently, there is little incentive to improve the quality of evidence (Hayes et al. 2013).

Although many commentators have discussed the technicalities of biomarker study design that contribute to these problems,[1] some of the more fundamental philosophical issues remain unexplored. For example, in contrast to traditional clinical development, for which the unit of testing is a single agent (e.g., a drug), the relevant unit of testing in biomarker development is a more complex "intervention ensemble" (Hey and Kimmelman 2014). This ensemble combines (1) a diagnostic assay protocol, (2) a target biomarker, and (3) a particular course of therapy. Successfully translating this three-part "biomarker ensemble" into the clinic is a more challenging task than standard drug development, since it requires that researchers determine the optimal parameters and procedures for all of these components.

In this paper, I will argue that this more complex ontology of testing gives rise to a particularly vexing problem of theoretical underdetermination. That is, the network of biological and pathophysiological theories motivating a biomarker experiment is often sufficiently complicated and uncertain that it frustrates valid interpretation of the experimental results. To illustrate this problem, I begin in the next section by discussing a case-study in biomarker diagnostic development: The anti-cancer therapy, temozolomide (TMZ), has been extensively tested in patients with glioblastoma multiforme (GBM) whose tumor specimens test positive for methylation of the O-6-methylguanine DNA methyltransferase (MGMT) promoter region of their DNA. But despite more than ten years of research, and seemingly wide agreement that a patient's MGMT-methylation status can be predictive of their response to TMZ, MGMT diagnostic testing is not yet implemented in the clinic.

In §3, I argue that the failure to fully translate MGMT testing into the clinical setting can be explained (at least in part) by underdetermination. Specifically, I show that many different theories are implicated in a biomarker experiment—these include the theory of how the drug interacts with the disease, the theory explaining the biomarker's relationship to the drug and disease, and the theory explaining how the diagnostic assay(s) can reliably determine a patient's biomarker status. In the case of MGMT testing for GBM response to TMZ, most of these theories are not yet well-confirmed (much less well-understood) and this uncertainty makes it difficult for researchers to judge the quality of emerging evidence.

[1] For example, the journal *Clinical Trials* (Oct. 2013) devoted an entire special issue to the technical challenges in designing and conducting biomarker studies.

In §4, I argue that overcoming this difficulty requires greater coordination across the biomarker research trajectory. Using the "Accumulating Evidence and Research Organization" (AERO) model—a graph-theoretic tool for representing and analyzing a scientific research program (Hey et al. 2013)—I then sketch an account of how coordination across a series of biomarker experiments could be evaluated. I ultimately conclude that the quality of a biomarker experiment must be judged by the contribution it makes to a coordinated research effort.

2. *The Case-Study: GBM, TMZ, and MGMT*

GBM is a common and severe form of brain cancer. Most GBM patients will die within a year from the time of diagnosis. Prior to TMZ, the standard of care was surgical resection of the tumor (if feasible), radiotherapy, and treatment with a nitrosourea agent (e.g., carmustine, lomustine). Nitrosouras and TMZ are both alkylating agents, which inhibit the cancer cells' repair activity by attaching an alkyl group to its DNA. However, TMZ is from the triazene class of drugs, and derived specifically from another widely-tested (although quite toxic) cancer drug, dacarbazine.

In 2005, TMZ received FDA approval as a first-line therapy against GBM. This decision was largely based on the results of a single clinical trial conducted by the European Organization for Research and Treatment of Cancer (EORTC) (Stupp et al. 2005). The EORTC trial included 287 patients and compared TMZ plus radiotherapy treatment versus radiotherapy treatment alone. They observed an improvement in the median survival of patients by 2.5 months (14.6 month survival in the TMZ + radiotherapy arm versus 12.1 month survival in the radiotherapy-only arm). However, almost half (49%) of the patients in the TMZ arm experienced "one or more serious or life-threatening adverse events," such as vomiting or convulsions (Pazdur 2013). Thus, at the time of TMZ's FDA licensure, the best a GBM patient prescribed TMZ could hope for was about 2-3 more months of survival, albeit with relatively high risk of life-threatening side-effects. Unfortunately, for the majority of GBM patients who still do not respond to TMZ therapy, they are just as likely to experience the serious side-effects without any survival benefit.

This is precisely the kind of situation that motivates research into personalized medicines. Despite the fact that TMZ is an improvement on the previous standard of care, only some GBM patients respond to the chemotherapy. The hope for personalized medicine is that there is a testable biological property that distinguishes the responding from the non-responding tumors. If this property can be prospectively identified, then it can improve the balance of risk and benefit for all patients: responders get effective treatment; non-responders are spared increased morbidity.

Fortunately, a subsequent retrospective analysis of the EORTC trial data by Hegi et al. (2005) revealed a biomarker—methylation of the MGMT promoter region—that seemed to identify a larger portion of the TMZ-responsive patients. MGMT is a gene that encodes for a DNA repair protein that fixes the damage caused by alkylating agents. In theory, if the promoter region of the MGMT gene is methylated, then this repair activity should be inhibited, and therefore, the patient should respond better to alkylating agents like TMZ (Esteller et al. 2000). Hegi et al. (2005) were able to show that for the patients given TMZ in the EORTC trial, the median survival for those with methylated MGMT was 21.7 months, versus 12.7 months for patients with unmethylated MGMT.

This 9-month difference in median survival essentially launched the MGMT biomarker research program. In the nine years since Hegi et al.'s study, the hypothesis that methylated MGMT should predict TMZ response has been tested in over three dozen studies. Yet, despite relatively wide agreement that MGMT's status has the capacity to predict TMZ response, MGMT testing is still not recommended for general clinical use. For example, the treatment guidelines of The National Comprehensive Cancer Network still only recommends MGMT diagnostic testing for very elderly (ages 70 and up) GBM patients (NCCN et al. 2012). The rationale is that these elderly patients are more likely to succumb to the toxicity of TMZ, whether they respond to therapy or not. And although this kind of stratification by age is certainly an advance in quality of care for the elderly subpopulation, it is still a long way from the more ambitious goal of personalized medicine—i.e., a predictive MGMT diagnostic that can pick out all (or most) of the patient population who will benefit from TMZ.

3. *Underdetermination and the Biomarker Ensemble*

The contrast between the successful translation of TMZ as a stand-alone agent and its (as yet) unsuccessful translation as a personalized medicine is instructive for understanding the epistemological differences between traditional drug development and biomarker diagnostic development. In traditional drug development, the unit of clinical translation is a single agent. Accordingly, we can think of the drug development program as largely driven by a single theory of disease and treatment mechanism, call it "T_D". In each clinical trial, there is a single hypothesis, H_D, about the effectiveness of the experimental treatment for a given patient population, *P*. If the new drug, "*D*" (developed on the basis of the theory T_D), improves outcomes for patients in *P*, then this is taken to confirm both H_D and T_D. Whereas, if *D* fails to improve outcomes, this is taken to disconfirm H_D, and it may also signal a problem with the theory T_D.

Yet, either outcome could also be erroneous—that is, an artifact of the experimental design or conduct, such as a protocol violation or a lack of control for bias. As many philosophers of science and medicine have now recognized, clinical development is thus regularly encountering the problem of underdetermination (Anderson 2006, Howick 2009, Hey and Weijer 2013, Chin-Yee 2014, Hey 2015).[2] The basic epistemic concern is that whether the experiment succeeds or fails—whether a clinical trial has a positive or negative result—there is always some uncertainty about the reliability of the result, the tested hypothesis, and the truth of the underlying theories motivating the experiment. Therefore, scientific judgment is needed to draw out the appropriate epistemic implications for T_D and H_D (Hey and Weijer 2013).

Now applying this formalization to the TMZ case, we can understand the aforementioned EORTC trial as a test of the hypothesis, H_{TMZ}, that TMZ would be an effective

[2] I should acknowledge that there are a variety of formulations for the problem of underdetermination in the philosophical literature. It is also sometimes called "Duhem's problem" or the "Duhem-Quine Problem" (Ariew 1984, Needham 2000, Darling 2002, Howson and Urbach 2006, Worrall 2010). Since a discussion of the relative philosophical merits of these accounts is outside the scope of this essay, I will just state that I am largely following the logical formalization of the problem implicit in Worrall (2010) and explicit in Hey (2014).

treatment for adult GBM. Its positive result was taken by the FDA to confirm H_{TMZ} sufficiently for licensing the drug as a first-line treatment for the patient population, P_{GBM}. This result could also be taken to have provided further confirmation of the theory, T_{Alk}, which explains the general effectiveness of alkylating chemotherapies.

But as the problem of underdetermination teaches, the FDA's decision to approve TMZ for this indication was not straightforwardly entailed by the evidence. The decision also reflects a judgment about the methodological quality of the EORTC study and the clinical value of a 2.5 month improvement in patient survival, despite an increase in toxicity. Which is to say, the FDA could have ruled differently without any violation of reason. For example, they might have withheld approval on the grounds that the sample population in the EORTC trial was not representative of the larger GBM population (i.e., calling into question the trial's external validity). They could therefore have demanded another study be conducted with a different sample population. Or they might have argued that TMZ's safety profile was too poor, and that even though the theory T_{Alk} had been confirmed in the study, the relevant hypothesis, H_{TMZ}, had not yet been sufficiently confirmed, since the balance of risk and benefit was not positive enough to conclude *clinical effectiveness*, even if it had shown *biological efficacy* (Tunis et al. 2003).

However, I do not want to get bogged down here in contemplating counter-factual FDA judgments. The essential points is, first, to show how the problem of underdetermination can play out in the standard model of clinical translation; and second, to observe that this model of underdetermination does not accurately characterize biomarker development. This is because the relevant unit of translation is not a drug or a diagnostic per se, but an "intervention ensemble," which includes a sensitive and specific assay protocol, a rigorously defined biomarker-positive population, a particular drug dose and schedule, various co-interventions, delivery techniques, and so on (Hey and Kimmelman 2014). Successful biomarker translation requires that optimal (or clinically sufficient) values for each of these components have been determined.

To illustrate: Consider that Hegi et al.'s (2005) retrospective study used a methylation-specific polymerase chain reaction (MS-PCR) assay to amplify the MGMT promoter region of the patient's tumor specimen. The association they demonstrated between a methylated MGMT promoter and benefit from TMZ therefore depends on at least four other dimensions of epistemic uncertainty. These are:

1. The theory of the biomarker, call it "T_{MGMT}," explaining MGMT's relationship to the drug and disease (i.e., reduced activity of the DNA repair enzyme that counteracts alkylation therapy should correlate with increased drug activity);
2. The biomarker hypothesis, H_{MGMT}, that methylated MGMT will accurately distinguish the TMZ-responsive patients in the sample population, p_{GBM}, from the non-responsive;
3. The theory of the diagnostic, call it "T_{PCR}," explaining the assay's relationship to the marker, drug, and disease (i.e., that there exists an MS-PCR protocol that is sufficient to accurately detect the relevant region and extent of MGMT methylation in specimens drawn from P_{GBM});
4. The diagnostic hypothesis, H_{PCR}, that patient specimens testing positive for MGMT methylation with a specific MS-PCR protocol can reliably distinguish TMZ-responsive patients in p_{GBM} from the non-responsive.

The epistemic status for all of these elements necessarily mediates the interpretation of the study results. For example: If we observe a positive association between biomarker-status and treatment response, is this due to a truly predictive biomarker (i.e., confirmation of T_{MGMT}) or is it merely an artifact of the assay? Or if we observe a null result, does this indicate a useless biomarker (i.e., disconfirmation of T_{MGMT}) or might the biomarker still be predictive for a more narrowly defined patient population (i.e., disconfirmation of H_{MGMT}, but not necessarily disconfirmation of T_{MGMT})?

What if one kind of assay confirms a biomarker's utility and a second assay disconfirms it? If one assay is not already known to be more reliable than the other, then what does this discordance tell us about the relationship between the assays, the biomarker, and the treatment? Or perhaps the observed association between biomarker and response may be confounded by another biomarker altogether?

All of these questions bear on the judgments of experimental quality and the interpretation of evidence. Yet, insofar as any of the theories and hypotheses underlying the disease, drug, biomarker, or assay remain uncertain, it should accordingly reduce our confidence in providing definitive answers. Until researchers can assert with confidence that, for example, "*this* MS-PC assay protocol is 90% sensitive and specific for detecting methylated MGMT in the sample population, and testing positive confers ten-fold greater odds of responding to TMZ," then the final goal of personalized medicine has not yet been achieved.

Moreover, we should observe that biomarker experiments cannot support valid inferences independent of the basic scientific theories in the way that traditional, late-phase drug trials can. For example, Ashcroft (2004) argues that randomized controlled trials (RCTs) can provide valid answers to the clinical question of interest—i.e., "Should we prescribe drug A or B for the population with the condition P_C?"—without needing a true causal theory, T_A or T_B, for why the drugs are effective. Howick (2011) also makes a similar claim— arguing that the value of theoretical or mechanistic reasoning is overemphasized in medical research. In essence, these authors both suggest that the driving theories in an RCT may be entirely wrong and, at least in principle, it should make no difference at all to clinical practice. All that matters is for the RCT to get it right about which intervention is better (and that neither falls below the standard of competent medical care).[3]

However, biomarker development is not like this. To begin with, most biomarker experiments are retrospective analyses of archived tissues and are not conducted in the context of an RCT. Therefore, they are not directly addressing a pragmatic, clinical question about which biomarkers we should use in practice. But even more fundamentally, biomarker development is explicitly trying to leverage the theories and techniques of basic science toward better treatment practices. Therefore, by definition, biomarker translation cannot be successful until it has resolved these theoretical uncertainties.

The development of another personalized medicine, the epidermal growth factor receptor (EGFR)-inhibitor, cetuximab, provides an illuminating example here: Cetuximab was initially approved by the FDA for use against metastatic colorectal cancer for patients whose tumors tested positive for elevated EGFR-protein expression. However,

[3] I should note that the alleged independence of theory from RCTs is a controversial position. For example, Giacomini (2009), Hey (2014), and Kimmelman and London (2015) all argue that theories and theoretical understanding have an essential role to play in clinical development.

it was later discovered that non-EGFR-expressing patients also responded, and in fact, it was a different genetic marker altogether—mutation of the KRAS gene—that predicted response to cetuximab (Chung et al. 2005). Although cetuximab is generally considered a success for personalized medicine, this initial misidentification of the biomarker ensemble had serious repercussions for the quality of health care. Many patients who could have benefited from cetuximab (i.e., those with low EGFR-expression) were deprived of effective care; and patients who were not likely to respond (i.e., those with high EGFR-expression but KRAS-mutations) were prescribed an expensive, highly-toxic, and ineffective treatment.

These unfortunate consequences can be attributed to an inadequate understanding and testing of the basic theories concerning cetuximab, EGFR-inhibition, and protein expression assays. Since cetuximab was explicitly developed as an EGFR-pathway inhibitor, it is certainly plausible to think it should work better for patients whose tumors showed elevated EGFR protein expression. However, researchers did not actually test this theory before the biomarker ensemble was approved. Indeed, the trials in support of cetuximab's approval as a personalized medicine systematically excluded patients with low EGFR-expression (Cunningham et al. 2004). This means that although researchers were able to show and explain improvements over the standard of care on the basis of a flawed theory, they left a critical dimension of uncertainty unresolved. In retrospect, it is clear that this lack of theoretical resolution resulted in worse patient outcomes and greater health-care inefficiencies—exactly the opposite of what personalized medicine is supposed to achieve.

In the case of TMZ and MGMT testing, at least some of these dimensions of theoretical uncertainty do seem to be resolved. The general characterization of the relevant biomarker and patient population parameters are largely settled (MGMT and newly-diagnosed GBM, respectively). But uncertainty still remains about the optimal assay (Febbo et al. 2011, Wick et al. 2014). Most studies examining an MGMT biomarker for predicting TMZ response have followed Hegi et al. (2005) in using MS-PCR. But some critics argue that this assay is unreliable when used with archived tissue samples (Dunn et al. 2009). Given that many MGMT biomarker studies are retrospective analyses of archived tissues (including Hegi et al.'s study), this criticism calls into question the quality of much of the available evidence.

One of the alternative assays to MS-PCR is immunohistochemistry (IHC). Rather than amplifying the promoter region of the DNA, IHC assays expression of the MGMT repair protein directly. Although IHC also appears to predict TMZ-responders (Chinot et al. 2007), studies that have directly compared IHC and MS-PCR have found that the methods are not interchangeable (Mellai et al. 2009, Lechapt-Zalcman et al. 2012). On its face, this discordance suggests problems with the underlying theories T_{MGMT}, T_{PCR}, and T_{IHC}. Specifically, it challenges the assumed identity and mechanistic relationship between MGMT promoter methylation (as assayed by MS-PCR) and MGMT protein expression (as assayed by IHC). Researchers have generally assumed that both properties are causally connected to alkylation-damage repair, and should therefore predict response to TMZ, but this theoretical uncertainty has not been resolved.

To complicate matters further, Lechapt-Zalcman et al. (2012) found that although MS-PCR and IHC selected different biomarker-positive patient populations in their sample, the hazard ratio for death was nearly identical (0.46 for MS-PCR; 0.45 for IHC). From a health-care system perspective, this suggests that either assay may be suitable for clinical use, since each selects a population that is more likely to benefit from TMZ therapy. However, from a patient perspective, discordance between the assays is a cause for concern. A patient classified as MGMT-negative on one test may be MGMT-positive on the other. Given that positivity on either test correlates to better outcomes, if treatment decisions are to be based on the test result, it would be in the patient's best interest to demand both tests. But even then, the ever-present possibility of misclassification (that is, either a false-positive or false-negative result) might encourage patients to skip the added cost and burden of the diagnostics altogether and simply go ahead and prescribe TMZ.[4]

This lingering uncertainty about the theories underlying alternative MGMT assays, as well as the possibility of misclassification, helps to explain why MGMT testing is not yet implemented in the clinic. In essence, the unit of translation has not yet been determined. Or in other words: The outstanding question that needs to be answered in future biomarker experiments is not simply if MGMT can be predictive of response to TMZ (since we already know that it can). Rather, the question now is: "What are the necessary and sufficient conditions for leveraging a methylated-MGMT diagnostic in order to maximize the therapeutic benefit of first-line TMZ in adult GBM patients?" To adequately address this question—and avoid the kinds of mistakes made with cetuximab—researchers will need to resolve the many remaining dimensions of uncertainty concerning the assays (e.g., T_{PCR}, T_{IHC}, H_{PCR}, H_{IHC},...).

4. Study Quality and the Division of Epistemic Labor

One obvious recommendation for resolving these uncertainties is to design and conduct more rigorous experiments. As I noted in the introduction, the majority of criticisms in the biomarker literature have focused on this aspect of the problem. For example, Simon et al. (2009) point out that many biomarker studies do not report how (or if) they blinded assay evaluators to the patient's clinical status. Study reports also often fail to provide a prospective definition for the cut-off value used when judging biomarker-status and do not adequately describe the use of control specimens.

For retrospective analyses, which typically rely on "convenience samples" of archived specimens from a local institutional biobank, Simon et al. also emphasize the danger of biased sampling. The worry is that there may be a causally relevant difference between archived specimens that give meaningful assay results from those deemed "inadequate" for testing. Thus, they argue that the sample patient population, from which all specimens will be drawn, must be prospectively defined with clear eligibility criteria. And then at least 66% of the defined patient population must provide specimens that produce meaningful assay results.

It is also rare for biomarker studies to report estimates of the sensitivity and specificity for their assays. And yet, without this information, it is difficult to judge whether the assay used in a study has adequately demarcated the biomarker-positive and biomarker-negative populations. This uncertainty weakens any inferences about the biomarker's clinical validity and utility (McShane and Polley 2013). As noted above, for studies that use multiple

[4] According to Wick et al. (2014), this is precisely what physicians and patients are currently doing.

assays, disagreement in classification between the assays is potentially informative for understanding the posited theoretical relationships among the biomarker, the assays, and the disease. But of course, if misclassification rates are not reported, then the opportunity to reduce this uncertainty is lost.

Some of these technical problems could be straightforwardly addressed through better reporting practices. Indeed, this is the goal of the REMARK guidelines, which provide a checklist of methodological items that every published biomarker experiment should report (McShane et al. 2005). Although adherence to these guidelines has been poor (Mallett et al. 2009), better reporting would be a big first step toward improving the general quality of evidence in the biomarker research enterprise.

However, we should be careful not to conflate the value of better reporting with the demand for more rigorous biomarker study designs. High-quality reporting is a necessary condition for supporting sound judgments about the evidence in the published literature (Glasziou et al. 2014). We can therefore assert the need for better reporting without qualification. In contrast, the call for more rigorous biomarker study designs needs to be carefully qualified. This is due to the division of epistemic labor in medical research between exploratory and confirmatory investigations (Kimmelman et al. 2014, Hey and Kimmelman 2014).

In the biomarker context, we can think of exploratory studies as hypothesis-generating and hypothesis-pruning. Their role is to expand or narrow the parameter space for a clinically viable intervention ensemble. Whereas confirmatory studies are hypothesis-testing. Their role is to put a particular set of ensemble parameters to a decisive test. Accordingly, it would be overhasty to dismiss some of the less rigorous, more exploratory biomarker studies as uninformative or misleading. For example, rather than prospectively defining a biomarker cut-off, an exploratory study might only retrospectively compare the predictive utility for multiple biomarker cut-offs in their study population. Or it might compare multiple different assays, and therefore, the rate of successful results (if agreement between assays is a prespecified requirement for success) may fall below the 66% threshold. Although these kinds of investigations should not be taken to provide definitive evidence for or against a biomarker's utility, they can still make a valuable epistemic contribution. But the precise value of this contribution will be relative to the accumulating state of scientific evidence. In the early stages of development, it should not be troubling to find less rigorous and more exploratory studies, since investigators may still be foraging for viable ensemble parameters. However, as more evidence accumulates about the range of viable biomarker ensemble parameters, we should expect that exploratory studies will give way to more confirmatory investigations.

This model for the division of epistemic labor also implies that it would be inefficient to demand that all biomarker studies have a rigorous, confirmatory orientation. As important as confirmatory studies are for providing definitive answers, a premature confirmatory investigation—i.e., one conducted before researchers can be confident that they have adequately explored the parameter space—would only show that one particular set of ensemble parameters is not useful. When the range of viable ensemble parameters is poorly understood, this is not particularly valuable information. Thus, the ideal is to use both exploratory and confirmatory orientations in a complimentary and coordinated research effort. Exploratory studies map the ensemble parameter space in search of the most promising biomarker ensemble. Confirmatory studies then take this ensemble and put it to a rigorous test.

5. *Robustness and Coordination of Research Activities*

In a recent article, I argued that the entire process of clinical drug development can be thought of as a multi-modal robustness analysis (Hey 2015). In virtue of its translational trajectory through in vitro, in vivo, and human experiments, an approved drug is a robust intervention. Much the same can be said for biomarker diagnostic development, except, as we saw in §3, there are more parameters of uncertainty to resolve in translating a biomarker ensemble. A truly successful biomarker diagnostic has been analytically validated, then shown to predict patient benefit in archived specimens, and finally, passed decisive evaluation in clinical trials (Simon 2012). Indeed, success through this trajectory should be sufficient to furnish physicians with a robust understanding about how to use and interpret the results of a diagnostic test.

How then might a series of biomarker experiments coordinate their activities toward this robust clinical understanding? As a first step, we can clarify some kinds and properties of a biomarker's translational trajectories that we do not want. First, we do not want a "black box" translation—that is, a trajectory which identifies an effective biomarker ensemble, but fails to either adequately explore the parameter space or illuminate anything about the underlying theories. Physicians will often need to make adjustments to an intervention ensemble in practice—e.g., reducing dose in the face of patient toxicity or deciding about treatment in cases of borderline diagnosis. If physicians do not have any empirical or theoretical basis for making these judgments, then the research system has failed to adequately inform clinical practice. And as the cetuximab case makes clear, having a "black box" personalized therapy can actually diminish the overall quality of health care.

Second, we should strive for efficiency of translation, avoiding unnecessary duplication of study questions. The pool of human and material resources for research is limited. Some replication or verification of study results is certainly desirable, but once a question has been answered, it is important for investigators to articulate novel—and still unsettled—hypotheses. Yet, as noted above, efficiency must be balanced against the risk of a premature confirmatory investigation. A biomarker hypothesis that has only been tested in exploratory-type studies is not yet ready to be folded into a large RCT. Positive results should first be replicated and then refined through a systematic robustness analysis that identifies the set of assay, biomarker, therapy, and population parameters most likely to prove clinically useful (McShane and Polley 2013). Once this process is complete, and the biomarker ensemble is sufficiently defined, only then is it truly ready for a decisive clinical evaluation in an RCT.

Third, we want to avoid unnecessary risk. This is related to the second concern, since the risks of study participation are not sufficiently redeemed in either a study that investigates a settled research question or a study that inefficiently expends research resources on a premature confirmatory test. Safety information is also accumulating over the course of the research trajectory, and this information should be used to design safer human experiments. Thus, prospective biomarker studies—particularly RCTs that allocate patient-subjects on the basis of their biomarker status—that are not designed on the basis of the most recent available evidence may be exposing patients to unnecessary risks and burdens.

These three restrictions on the ideal biomarker translation are grounded in the principles of research ethics—specifically, the principles of risk minimization, social value, and protecting the integrity of the research enterprise. These principles also imply two other properties of the ideal translation that are worth making explicit: First, every study should be explicitly designed to incorporate and build upon the accumulating state of evidence. This requirement not only helps to satisfy the concerns about unnecessary duplication and risk, it also facilitates a more efficient and robust research process. Second, the testing hypotheses should evolve over time. When a biomarker is first identified, it is common to see vague and qualitative hypotheses about its role in predicting treatment response (e.g., "MGMT repairs the damage caused by TMZ, therefore we investigated the impact of methylated MGMT on treatment response"). These kinds of qualitative claims are sufficient early on in the research program, but as evidence accumulates, the biomarker hypotheses should become more precise and quantitative (e.g., "We expect patients that test positive for methylated-MGMT using MS-PCR to average 5-months longer survival"). Confirmation or refutation of these quantitative claims is needed to support the implementation of biomarker diagnostics in clinical practice. Physicians, in particular, need this information in order to effectively interpret the results of diagnostics for their patients. Therefore, it is incumbent on researchers to progressively refine the viable hypotheses.

Putting all of these principles together, we can articulate a heuristic research strategy for biomarker translation: Ideally, we should like to see early studies exploring a wide array of ensemble parameter values—that is, testing multiple biomarkers, comparing the accuracy of multiple assays, evaluating the biomarker's utility for different patient populations, etc. Once some promising ensemble combinations have been identified, later studies submit these to a definitive test.

Yet it is one thing to articulate a coordinated research strategy, it is quite another to implement or evaluate it. Indeed, the need for greater coordination has been recognized across the clinical research enterprise—not just in the realm of biomarker development (Boucher et al. 2009, Seymour et al. 2010, Hayes et al. 2013, Hay et al. 2014). But, as Hey et al. (2013) observe, one of the barriers to improving coordination has been a lack of tools or frameworks for analyzing or communicating the state of coordination.

As a potential remedy to this problem, Hey et al. (2013) developed the "AERO model"—a graph-theoretic model for representing the accumulating state of evidence and organization of research activities within a given scientific domain. Briefly, this model visually represents a set of study reports (ideally drawn from a systematic literature review) as nodes in a directed network, arranged along the x-axis by time, and stratified on the y-axis by various study parameters. Arrows (or edges) between nodes are then used to indicate inferential relationships between studies.

Table 1 lists 37 studies that investigated the utility of an MGMT assay for predicting TMZ response in adult GBM patients.[5] In addition to basic study details—e.g., authors, year of publication, retrospective versus prospective design, type of assay, and significance of outcome—this table also includes a rough quality metric, or "Q-score". This Q-score, based on the recommendations in Simon et al. (2009), is a 5-point scale, where 1 point each is awarded for reporting (1) blinded assessment; (2) prespecified definition for biomarker-positive status; (3) use of control specimens; and (4) biomarker status successfully determined for greater than 66% of collected specimens.

[5] This set of 37 studies is the result of a systematic review, whose technical details are described in Hey et al. (n.d.).

Table 1. Studies evaluating an MGMT assay for predicting first-line TMZ response in adult GBM. The "ID" column corresponds to the node in the AERO diagrams. "Q" is a 5-point quality score, where 1 point each is awarded for 1 point each is awarded for reporting (1) blinded assessment; (2) prespecified definition for biomarker-positive status; (3) use of control specimens; and (4) biomarker status successfully determined for greater than 66% of collected specimens. "OS-18" is the significance of the association between MGMT-status and 18-month patient survival.

ID	Authors	Year	Design	Assay(s)	Q	OS-18
01	Hegi et al.	2004	Phase 2	MS-PCR	2	Positive
02	Hegi et al.	2005	Retro	MS-PCR	2	Positive
03	Herrlinger et al.	2006	Phase 2	MS-PCR	2	Positive
04	Chinot et al.	2007	Phase 2	IHC	3	Null
05	Brandes et al.	2008	Retro	MS-PCR	1	Positive
06	Brandes et al.	2009	Phase 2	MS-PCR	1	Positive
07	Clarke et al.	2009	Phase 2	MS-PCR	1	Positive
08	Dunn et al.	2009	Retro	MS-PCR, PSQ	3	Positive
09	Mellai et al.	2009	Retro	MS-PCR, IHC, WB	3	Null
10	Panet-Raymond et al.	2009	Retro	MS-PCR	2	Null
11	Weller et al.	2009	Retro	MS-PCR	2	Positive
12	Ang et al.	2010	Retro	MS-PCR	2	Positive
13	Costa et al.	2010	Retro	MS-PCR	3	Null
14	Weiler et al.	2010	Phase 2	MS-PCR	3	Positive
15	Balana et al.	2011	Retro	MPS, IHC	4	Null
16	Chan et al.	2011	Retro	MS-PCR	1	Positive
17	Motomura et al.	2011	Retro	PSQ	3	Positive
18	Park et al.	2011	Retro	MS-PCR, IHC, MLPA	3	Null
19	Shah et al.	2011	Retro	MS-PCR, IHC, MLPA, et al.	3	Positive
20	Thon et al.	2011	Retro	MS-PCR	3	Positive
21	Watanabe et al.	2011	Retro	IHC	3	Null
22	Christians et al.	2012	Retro	MS-PCR, PSQ, MLPA	3	Positive
23	Iliadis et al.	2012	Retro	IHC, MLPA	3	Null
24	Karim et al.	2012	Phase 2	MS-PCR, IHC	3	Positive
25	Kim et al.	2012	Retro	MS-PCR	3	Null
26	Lam and Chambers	2012	Retro	MS-PCR	0	Positive
27	Lechapt-Zalcman et al.	2012	Retro	MS-PCR, IHC	4	Null
28	Quillien et al.	2012	Retro	MS-PCR, IHC, PSQ, et al.	3	Positive
29	Reifenberger et al.	2012	Retro	MS-PCR, PSQ	2	Null
30	Abhinav et al.	2013	Retro	MS-PCR	0	Null
31	Gilbert et al.	2013	Phase 3	MS-PCR	1	Positive
32	Gutenberg et al.	2013	Retro	MS-PCR	2	Null
33	Hsu et al.	2013	Retro	MS-PCR, IHC	4	Null
34	Lalezari et al.	2013	Retro	MS-PCR, IHC, BiSEQ	3	Positive
35	McDonald et al.	2013	Retro	MS-PCR, PSQ	3	Null
36	Romano et al.	2013	Retro	MS-PCR	2	Null
37	Sunwoo et al.	2013	Retro	MS-PCR, MLPA	2	Null

Figure 1 illustrates the AERO model applied to these studies. Circular nodes in the figure represent prospective clinical trials that included a biomarker component (the larger node is a phase 3 trial, the others are phase 2), square nodes represent retrospective studies. Nodes are white if the authors showed a statistically significant association between MGMT status and survival at 18 months; nodes are shaded if the association was not significant (i.e., null results).[6] The nodes are stratified here by the type(s) of diagnostic assay used, either MS-PCR (the most common), direct comparison of "Multiple" assays, IHC, or pyrosequencing (PSQ).

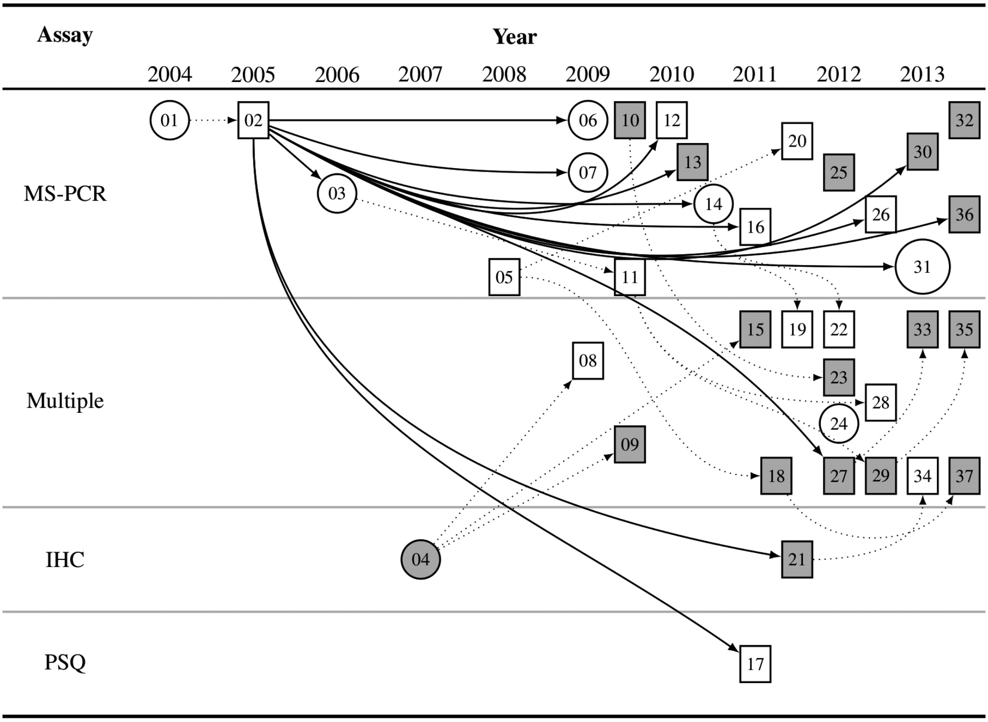

Figure 1. AERO graph for studies investigating MGMT testing as a predictive diagnostic for first-line TMZ therapy in adult GBM patients. Solid arrows are all references to Hegi et al. (2005)—the landmark retrospective study

The arrows between the nodes represent the most recent citation to other studies in the sample that appeared in either the introduction or methods section of the published report. For example, the arrow between node 2 and node 6 indicates that study 2 was the

[6] What is a clinically relevant outcome varies among types of cancer. As described earlier, median survival for a GBM patient given TMZ (without the use of a biomarker diagnostic) is around 14.6 months. Therefore, 18-month survival represents a clinically significant improvement that could justify the use of a predictive diagnostic.

most recently cited study in the introduction and methods section of study 6. While this method for representing the flow of evidence is certainly an imperfect proxy measure for coordination of research activity, it is still informative. The introduction and methods are those sections of a scientific report responsible for articulating the rationale for the tested hypothesis. Authors that do not explicitly reference recent and available evidence as the justification for their hypothesis are arguably conducting inefficient and ethically questionable research.

Figure 1 reveals a number of interesting properties of the TMZ-MGMT research program: First, the majority of studies were positive—finding a significant association between methylated-MGMT and the odds of survival to 18 months. Yet, the proportion of positive-to-null studies is not overwhelming. After 2011, this ratio actually shifts in favor of null results. Nevertheless, the only phase 3 trial in this sample did find a significantly positive association (Gilbert et al. 2013). This is a promising sign for the clinical utility of MGMT testing.

Second, we can observe that after 2011, more studies are directly testing multiple assays. This indicates an evolution of the testing hypothesis. Specifically, it reflects a shift in focus from the question "Is MGMT a predictive marker?" to "What is the optimal assay for maximizing the clinical utility of an MGMT marker?" However, this shift in the testing hypothesis also calls into question the epistemic value of some of the later studies that only looked at MS-PCR. Given the early positive results in this stratum, it is not surprising to see continued interest. And indeed, some this work is surely consistent with the need to validate early exploratory results. Nevertheless, by 2010 or 2011, when 10 or more studies have already retrospectively investigated this method, further investigations seem problematic. Why has this question not been adequately answered? What more can these studies plausibly teach us about the predictive value of an MS-PCR diagnostic?

This leads to a third point: The continued references to Hegi et al. (2005) as justification reflect a failure (on the part of some studies) to update their hypothesis in light of the accumulating state of evidence. 14 of the 30 total justifications are to this one study. While this is not troubling in the early years of the research program, as more studies are completed and published, it becomes ethically and scientifically problematic to ignore the growing body of evidence. This study's modest Q-score also undermines the explanation that it is so heavily cited because of its methodological rigor.

Finally, the overall "shape" of the trajectory appears to be the reverse of our ideal. That is, the wide-exploration of assay parameters—evident in the cluster of studies in the "Multiple" stratum post-2011—takes place later in the research trajectory. These studies are doing essential epistemic work toward resolving the biomarker ensemble parameters. But ideally, this work would be immediately after the identification of a promising marker. Instead, what we see is a large portion of the research program that was content for some time to investigate and re-investigate the same set of ensemble parameters.

Figure 2 builds on these criticisms by applying two additional layers of analysis. Nodes are now shaded according to their Q-score, and edges styled according to the recency of the citation. For the latter, the rule was as follows: Cited justifications to evidence within the previous 3 years are solid edges; citations to evidence older than this are represented with dotted edges. Although 3 years may be an arbitrary cut-off, for retrospective studies—which are often studies of convenience and can be completed much more quickly than a prospective trial—this should still be more than enough time to expect that authors have read and folded the more recent and relevant literature into their design.

This alternative coding scheme makes the contrast in epistemic value even clearer: Post-2011, experiments in the MS-PCR only stratum were generally of lower quality and tended to ignore the accumulating body of evidence. Only 1 of the 8 studies in this region of the graph cited recent evidence as motivation. The rest either cited Hegi et al. (2005) or nothing at all. Whereas the studies comparing multiple assays tended to be of higher quality and justified their investigations on the basis of more recent developments. Given that discordance between assays is essential for theoretical resolution, this activity is encouraging: The studies directly testing the most pressing dimensions of uncertainty (i.e., T_{PCR}, T_{IHC}...) also tended to better report their methods.

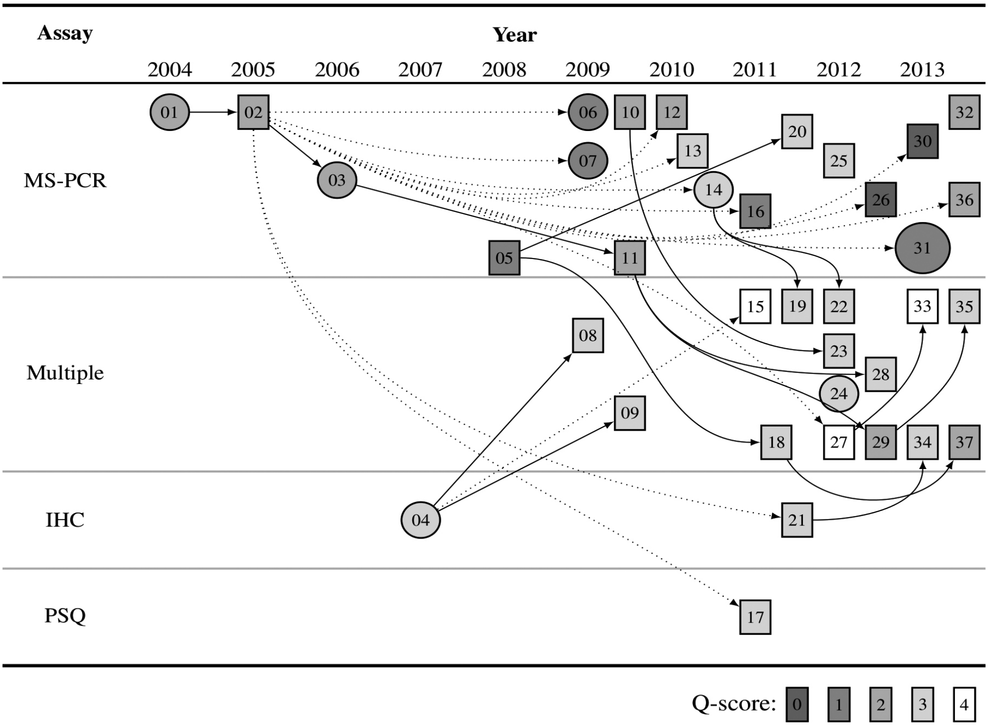

Figure 2. AERO graph re-coded to show quality and uptake of evidence. Nodes are shaded according to Q-score. Solid edges are citations to "recent" evidence in the sample (within 3 years of publication). Dotted edges are citations to "old" evidence (more than 3 years since publication)

Finally, we should note that all three of the studies with perfect Q-scores had null results. This is a less encouraging sign for MGMT testing. Yet, these all compared the same two assays: MS-PCR and IHC. Fortunately, these are not the only assays undergoing evaluation. Therefore, if we understand these three experiments as confirmatory (albeit retrospective) investigations, then the null result does not indicate that there is no clinically viable MGMT diagnostic. Instead, it indicates that MS-PCR and IHC, despite the promising preliminary evidence, may not be the most clinically useful diagnostics. This still leaves open the possibility that other, more useful assays may yet be found.

6. Conclusion

The case of MGMT diagnostic testing for GBM response to TMZ provides many philosophical lessons for understanding the dynamics of biomarker diagnostic translation and judging the quality of biomarker experiments. The most fundamental of these is that biomarker translation is a more complicated process than standard, single-agent translation. Biomarker experiments necessarily implicate multiple theories and hypotheses, and the uncertainty surrounding each of these must be addressed before a biomarker can be legitimately recommended for clinical use. While these theories are still in flux, it can be exceedingly difficult to judge the quality of emerging evidence—even more so given that each experiment may be testing a slight variation on the biomarker ensemble.

However, there is an optimistic lesson here: As the AERO model makes explicit, the trend of positive outcomes is consistent with a biomarker whose capacity to predict treatment response is widely accepted. The reporting quality and evidence uptake revealed in the "Multiple" stratum of figure 2 also highlights the fact that well-coordinated work is now being done where it is needed most: resolution of uncertainty concerning the optimal assay. Although MGMT testing is not yet recommended for general clinical use, this should give us reason to hope that underdetermination may soon be overcome and a clinically useful assay identified.

In sum, this case illustrates that although realizing the potential of personalized medicines is a high epistemic bar, it is not an insurmountable challenge. But meeting this challenge in an efficient way demands that researchers better coordinate their activities. At its root, the goal of personalized medicine is to furnish health-care providers with a robust understanding of the theories underlying a course of treatment: Why do only some patients respond to therapy? How can we prospectively test patients in order to distinguish the responders from the non-responders? Providing evidence-based answers to these questions requires a systematic exploration of the biomarker ensemble parameter space. This is the project of multiple experiments, each of which is responsible for incorporating the most recent available evidence into its study design. Therefore, what makes a high-quality biomarker experiment cannot solely be a function of its individual methods or rigor. It must be judged by its place in—and contribution to—this coordinated research effort.

REFERENCES

Abhinav, K., K. Aquilina, H. Gbejuade, M. La, K. Hopkins and V. Iyer. 2013. A Pilot Study of Glioblastoma Multiforme in Elderly Patients: Treatments, O-6-methylguanine-DNA Methyltransferase (MGMT) Methylation Status and Survival. *Clinical Neurology and Neurosurgery* 115 (8): 1375-1378.

Anderson, James A. 2006. The Ethics and Science of Placebo-controlled Trials: Assay Sensitivity and the Duhem-Quine Thesis. *Journal of Medicine and Philosophy* 31 (1): 65-81.

Ang, C., M.-C. Guiot, A. Ramanakumar, D. Roberge, and P. Kavan. 2010. Clinical Significance of Molecular Biomarkers in Glioblastoma. *The Canadian Journal of Neurological Sciences* 37 (5): 625-630.

Ariew, Roger. 1984. The Duhem Thesis. *The British Journal for the Philosophy of Science* 35 (4): 313-325.

Ashcroft, Richard E. 2004. Current Epistemological Problems in Evidence Based Medicine. *Journal of Medical Ethics* 30 (2): 131-135.

Balana, Carmen, Cristina Carrato, Jose Luis Ramirez, Andres Felipe Cardona, Mireia Berdiel, Jose Javier Sanchez, Miquel Taron et al. 2011. Tumour and Serum MGMT Promoter Methylation and Protein Expression in Glioblastoma Patients. *Clinical and Translational Oncology* 13 (9): 677-685.

Boucher, Helen W., George H. Talbot, John S. Bradley, John E. Edwards, David Gilbert, Louis B. Rice, Michael Scheld, Brad Spellberg, and John Bartlett. 2009. Bad Bugs, No Drugs: No Eskape! An Update from the Infectious Diseases Society of America. *Clinical Infectious Diseases* 48 (1): 1-12.

Brandes, Alba A., Enrico Franceschi, Alicia Tosoni, Francesca Benevento, Luciano Scopece, Valeria Mazzocchi, Antonella Bacci, Raffaele Agati, Fabio Calbucci, and Mario Ermani. 2009. Temozolomide Concomitant and Adjuvant to Radiotherapy in Elderly Patients with Glioblastoma. *Cancer* 115 (15): 3512-3518.

Brandes, Alba A., Enrico Franceschi, Alicia Tosoni, Valeria Blatt, Annalisa Pession, Giovanni Tallini, Roberta Bertorelle et al. 2008. MGMT Promoter Methylation Status Can Predict the Incidence and Outcome of Pseudoprogression After Concomitant Radiochemotherapy in Newly Diagnosed Glioblastoma Patients. *Journal of Clinical Oncology* 26 (13): 2192-2197.

Chan, D., M. Kam, B. Ma, S. Ng, J. Pang, C. Lau, D. Siu, B. Ng, X. Zhu, G. Chen, et al. 2011. Association of Molecular Marker O (6) Methylguanine DNA Methyltransferase and Concomitant Chemoradiotherapy with Survival in Southern Chinese Glioblastoma Patients. *Hong Kong Medical Journal= Xianggang yi xue za zhi/Hong Kong Academy of Medicine* 17 (3): 184-188.

Chin-Yee, Benjamin H. 2014. Underdetermination in Evidence-based Medicine. *Journal of evaluation in clinical practice* 20 (6): 921-927.

Chinot, Olivier L., Maryline Barrie, Stephane Fuentes, Nathalie Eudes, Sophie Lancelot, Philippe Metellus, Xavier Muracciole et al. 2007. Correlation Between O6-methylguanine-DNA Methyltransferase and Survival in Inoperable Newly Diagnosed Glioblastoma Patients Treated with Neoadjuvant Temozolomide. *Journal of Clinical Oncology* 25 (12): 1470-1475.

Christians, Arne, Christian Hartmann, Axel Benner, Jochen Meyer, Andreas von Deimling, Michael Weller, Wolfgang Wick, and Markus Weiler. 2012. Prognostic Value of Three Different Methods of MGMT Promoter Methylation Analysis in a Prospective Trial On Newly Diagnosed Glioblastoma. *PLoS One* 7 (3): e33449.

Chung, Ki Young, Jinru Shia, Nancy E Kemeny, Manish Shah, Gary K Schwartz, Archie Tse, Audrey Hamilton, Dorothy Pan, Deborah Schrag, Lawrence Schwartz, et al. 2005. Cetuximab Shows Activity in Colorectal Cancer Patients with Tumors That Do Not Express the Epidermal Growth Factor Receptor By Immunohistochemistry. *Journal of Clinical Oncology* 23 (9): 1803-1810.

Clarke, Jennifer L., Fabio M Iwamoto, Joohee Sul, Katherine Panageas, Andrew B Lassman, Lisa M DeAngelis, Adília Hormigo, Craig P Nolan, Igor Gavrilovic, Sasan Karimi, et al. 2009. Randomized Phase II Trial of Chemoradiotherapy Followed By Either Dose-dense or Metronomic Temozolomide for Newly Diagnosed Glioblastoma. *Journal of Clinical Oncology* 27 (23): 3861-3867.

Costa, Bruno M., Cl´audia Caeiro, Inês Guimarães, Olga Martinho, Teresa Jaraquemada, Isabel Augusto, L Osório, P Linhares, M Honavar, M Resende, et al. 2010. Prognostic Value of MGMT Promoter Methylation in Glioblastoma Patients Treated with Temozolomide-based Chemoradiation: A Portuguese Multicentre Study. *Oncology Reports* 23 (6): 1655-1662.

Cunningham, David, Yves Humblet, Salvatore Siena, David Khayat, Harry Bleiberg, Armando Santoro, Danny Bets, Matthias Mueser, Andreas Harstrick, Chris Verslype, et al. 2004. Cetuximab Monotherapy and Cetux-imab Plus Irinotecan in Irinotecan-refractory Metastatic Colorectal Cancer. *New England Journal of Medicine* 351 (4): 337-345.

Darling, Karen M. 2002. The Complete Duhemian Underdetermination Argument: Scientific Language and Practice. *Studies In History and Philosophy of Science Part A* 33 (3): 511-533.

Dunn, J., A. Baborie, F. Alam, K. Joyce, M. Moxham, R. Sibson, D. Crooks, D. Husband, A. Shenoy, A. Brodbelt et al. 2009. Extent of MGMT Promoter Methylation Correlates with Outcome in Glioblastomas Given Temozolomide and Radiotherapy. *British Journal of Cancer* 101 (1): 124-131.

Esteller, Manel, Jesus Garcia-Foncillas, Esther Andion, Steven N Goodman, Oscar F. Hidalgo, Vicente Vanaclocha, Stephen B. Baylin, and James G. Herman. 2000. Inactivation of the DNA-repair Gene MGMT and the Clinical Response of Gliomas to Alkylating Agents. *New England Journal of Medicine* 343 (19): 1350-1354.

Febbo, Phillip G., Marc Ladanyi, Kenneth D. Aldape, Angelo M. De Marzo, M. Elizabeth Hammond, Daniel F Hayes, A. John Iafrate, R. Kate Kelley, Guido Marcucci, Shuji Ogino, et al. 2011. NCCN Task Force Report: Eval-uating the Clinical Utility of Tumor Markers in Oncology. *Journal of the National Comprehensive Cancer Network* 9 (Suppl 5): S-1.

Giacomini, Mita. 2009. Theory-based Medicine and the Role of Evidence: Why the Emperor Needs New Clothes, Again. *Perspectives in Biology and Medicine* 52 (2): 234-251.

Gilbert, Mark R., Meihua Wang, Kenneth D. Aldape, Roger Stupp, Monika E. Hegi, Kurt A. Jaeckle, Terri S. Armstrong, Jeffrey S. Wefel, Minhee Won, Deborah T. Blumenthal et al. 2013. Dose-dense Temozolomide for Newly Diagnosed Glioblastoma: A Randomized Phase III Clinical Trial. *Journal of Clinical Oncology* 31 (32): 4085-4091.

Glasziou, Paul, Douglas G. Altman, Patrick Bossuyt, Isabelle Boutron, Mike Clarke, Steven Julious, Susan Michie, David Moher and Elizabeth Wager. 2014. Reducing Waste from Incomplete or Unusable Reports of Biomedical Research. *The Lancet* 383 (9913): 267-276.

Gutenberg, A., H. C. Bock, W. Brück, L. Doerner, H. M. Mehdorn, W. Roggendorf, M. Westphal, J. Felsberg, G. Reifenberger and A. Giese. 2013. MGMT Promoter Methylation Status and Prognosis of Patients with Primary or Recurrent Glioblastoma Treated with Carmustine Wafers. *British Journal of Neurosurgery* 27 (6): 772-778.

Hay, Michael, David W. Thomas, John L. Craighead, Celia Economides and Jesse Rosenthal. 2014. Clinical Development Success Rates for Investigational Drugs. *Nature Biotechnology* 32 (1): 40-51.

Hayes, Daniel F., Jef Allen, Carolyn Compton, Gary Gustavsen, Debra G. B. Leonard, Robert McCormack, Lee Newcomer, Kristin Pothier, David Ransohof, Richard L. Schilsky, Ellen Sigal, Sheila E. Taube and Sean R. Tunis. 2013. Breaking a Vicious Cycle. *Science Translational Medicine* 5 (196): 196cm6.

Hegi, Monika E., Annie-Claire Diserens, Sophie Godard, Pierre-Yves Dietrich, Luca Regli, Sandrine Ostermann, Philippe Otten, Guy Van Melle, Nicolas de Tribolet and Roger 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.

Hegi, Monika E., Annie-Claire Diserens, Thierry Gorlia, Marie-France Hamou, Nicolas de Tribolet, Michael Weller, Johan M. Kros, Johannes A. Hainfellner, Warren Mason, Luigi Mariani et al. 2005. MGMT Gene Silencing and Benefit from Temozolomide in Glioblastoma. *New England Journal of Medicine* 352 (10): 997-1003.

Herrlinger, Ulrich, Johannes Rieger, Dorothee Koch, Simon Loeser, Britta Blaschke, Rolf-Dieter Kortmann, Joachim P. Steinbach, Thomas Hundsberger, Wolfgang Wick, Richard Meyermann et al. 2006. Phase II Trial of Lomustine Plus Temozolomide Chemotherapy in Addition to Radiotherapy in Newly Diagnosed Glioblastoma: UKT-03. *Journal of Clinical Oncology* 24 (27): 4412-4417.

Hey, Spencer Phillips, 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.

Hey, Spencer Phillips and Charles Weijer. 2013. Assay Sensitivity and the Epistemic Contexts of Clinical Trials. *Perspectives in Biology and Medicine* 56 (1): 1-17.

Hey, Spencer Phillips. 2014. Theory Testing and Implication in Clinical Trials. In *Philosophy of Science Association 24th Biennial Meeting*.

— 2015. Robust and Discordant Evidence: Methodological Lessons from Clinical Research. *Philosophy of Science* 82 (1): 55-75.

Hey, Spencer Phillips and Jonathan Kimmelman. 2014. The Risk-Escalation Model: A Principled Design Strategy for Early-Phase Trials. *Kennedy Institute of Ethics Journal* 24 (2): 121-139.

Howick, J. 2009. Questioning the Methodologic Superiority of 'Placebo' Over 'Active' Controlled Trials. *The American Journal of Bioethics* 9: 34-48.

Howick, Jeremy. 2011. Exposing the Vanities —and A Qualified Defense— of Mechanistic Reasoning in Health Care Decision Making. *Philosophy of Science* 78 (5): 926-940.

Howson, Colin and Peter Urbach. 2006. *Scientific Reasoning: The Bayesian Approach*. Open Court Publishing.

Hsu, Chih-Yi, Shih-Chieh Lin, Hsiang-Ling Ho, Yi-Chun Chang-Chien, Sanford P-C Hsu, Yu-Shu Yen, Ming-Hsiung Chen, Wan-You Guo and Donald M-T Ho. 2013. Exclusion of Histiocytes/Endothelial Cells and Using Endothelial Cells as Internal Reference are Crucial for Interpretation of MGMT Immunohistochemistry in Glioblastoma. *American Journal of Surgical Pathology* 37 (2): 264-271.

Iliadis, Georgios, Vassiliki Kotoula, Athanasios Chatzisotiriou, Despina Televantou, Anastasia G. Eleftheraki, Sofia Lambaki, Despina Misailidou, Panagiotis Selviaridis and George Fountzilas. 2012. Volumetric and MGMT Parameters in Glioblastoma Patients: Survival Analysis. *BMC Cancer* 12 (1): 3.

Karim, Khaled Abdel, M. M. El Mahdy, M. M. Abdel Wahab, L. R. Ezz EI Arab, A. El Shehaby and S. Abdel Raouf. 2012. Temozolomide and Radiotherapy in Newly Diagnosed Glioblastoma Patients: O6-methylguanine-DNA Methyltransferase (MGMT) Promotor Methylation Status and Ki-67 as Biomarkers for Survival and Response to Treatment. *The ChineseGerman Journal of Clinical Oncology* 11 (3): 168-176.

Kim, Young Suk, Se Hoon Kim, Jaeho Cho, Jun Won Kim, Jong Hee Chang, Dong Suk Kim, Kyu Sung Lee and Chang-Ok Suh. 2012. MGMT Gene Promoter Methylation as a Potent Prognostic Factor in Glioblastoma Treated with Temozolomide-based Chemoradiotherapy: A Single-institution Study. *International Journal of Radiation Oncology*Biology*Physics* 84 (3): 661-667.

Kimmelman, Jonathan and Alex John London. 2015. The Structure of Clinical Translation: Efficiency, Information, and Ethics. *Hastings Center Report* 45: 1-7.

Kimmelman, Jonathan, Jeffrey S. Mogil and Ulrich Dirnagl. 2014. Distinguishing Between Exploratory and Confirmatory Preclinical Research Will Improve Translation. *PLoS Biology* 12 (5): e1001863.

Kyzas, Panayiotis A., Konstantinos T. Loizou and John P. A. Ioannidis. 2005. Selective Reporting Biases in Cancer Prognostic Factor Studies. *Journal of the National Cancer Institute* 97 (14): 1043-1055.

Lalezari, Shadi, Arthur P. Chou, Anh Tran, Orestes E. Solis, Negar Khanlou, Weidong Chen, Sichen Li, Jose A. Carrillo, Reshmi Chowdhury, Julia 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.

Lam, Nadine and Carole R. Chambers. 2012. Temozolomide Plus Radiotherapy for Glioblastoma in a Canadian Province: Efficacy Versus Effectiveness and the Impact of O6-methylguanine-DNA-methyltransferase Promoter Methylation. *Journal of Oncology Pharmacy Practice* 18 (2): 229-238.

Lechapt-Zalcman, Emmanu`ele, Gu´ena¨elle Levallet, Audrey Emmanuelle Dugu´e, Anne Vital, Marie-Dani`ele Diebold, Philippe Menei, Philippe Colin, Philippe Peruzzy, Evelyne Emery, Myriam 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.

Mallett, S., A. Timmer, W. Sauerbrei and D. G. Altman. 2009. Reporting of Prognostic Studies of Tumour Markers: A Review of Published Articles in Relation to REMARK Guidelines. *British Journal of Cancer* 102 (1): 173-180.

McDonald, K. L., R. W. Rapkins, J. Olivier, L. Zhao, K. Nozue, D. Lu, S. Tiwari, J. Kuroiwa-Trzmielina, J. Brewer, H. R. Wheeler et al. 2013. The T Genotype of the MGMT $C > T$ (rs16906252) Enhancer Single-nucleotide Polymorphism (SNP) Is Associated with Promoter Methylation and Longer Survival in Glioblastoma Patients. *European Journal of Cancer* 49 (2): 360-368.

McShane, Lisa M., Douglas G. Altman, Willi Sauerbrei, Sheila E. Taube, Massimo Gion, Gary M. Clark et al. 2005. Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK). *Journal of the National Cancer Institute* 97 (16): 1180-1184.

McShane, Lisa M. and Mei-Yin C. Polley. 2013. Development of Omics-based Clinical Tests for Prognosis and Therapy Selection: The Challenge of Achieving Statistical Robustness and Clinical Utility. *Clinical Trials*, 10 (5): 653-665.

Mellai, Marta, Valentina Caldera, Laura Annovazzi, Adriano Chi`o, Michele Lanotte, Paola Cassoni, Gaetano Finocchiaro and Davide Schiffer. 2009. MGMT Promoter Hypermethylation In a Series of 104 Glioblastomas. *Cancer Genomics-Proteomics* 6 (4): 219-227.

Motomura, Kazuya, Atsushi Natsume, Yugo Kishida, Hiroyuki Higashi, Yutaka Kondo, Yoko Nakasu, Tatsuya Abe, Hiroki Namba, Kenji Wakai and Toshihiko Wakabayashi. 2011. Benefits of Interferon-*β* and Temozolomide Combination Therapy for Newly Diagnosed Primary Glioblastoma with the Unmethylated MGMT Promoter. *Cancer* 117 (8): 1721-1730.

NCCN et al. 2012. National Comprehensive Cancer Network Clinical Practice Guidelines in Oncology: Central Nervous System Cancers. Version 2. 2013.

Needham, Paul. 2000. Duhem and Quine. *Dialectica* 54 (2): 109-132.

Panet-Raymond, Valerie, Luis Souhami, David Roberge, Petr Kavan, Lily Shakibnia, Thierry Muanza, Christine Lambert, Richard Leblanc, Rolando Del Maestro, Marie-Christine Guiot et al. 2009. Accelerated Hypofractionated Intensity-modulated Radiotherapy with Concurrent and Adjuvant Temozolomide for Patients with Glioblastoma Multiforme: A Safety and Efficacy Analysis. *International Journal of Radiation Oncology*Biology*Physics* 73 (2): 473-478.

Park, Chul-Kee, JinWook Kim, Su Youn Yim, Ah Reum Lee, Jung Ho Han, Chae-Yong Kim, Sung-Hye Park, Tae Min Kim, Se-Hoon Lee, Seung Hong Choi et al. 2011. Usefulness of MS-MLPA for Detection of MGMT Promoter Methylation in the Evaluation of Pseudoprogression in Glioblastoma Patients. *Neuro-oncology* 13 (2): 195-202.

Pazdur, Richard. 2013. FDA Approval for Temozolomide. Http://www.cancer.gov/cancertopics/druginfo/fda-temozolomide (retrieved August 7, 2014).

Quillien, Véronique, Audrey Lavenu, Lucie Karayan-Tapon, Catherine Carpentier, Marianne Labussière, Thierry Lesimple, Olivier Chinot, Michel Wager, Jérome Honnorat, Stephan Saikali et al. 2012. Comparative Assessment of 5 methods (Methylation-specific Polymerase Chain Reaction, Methylight, Pyrosequencing, Methylation-sensitive High-resolution Melting, and Immunohistochemistry) to Analyze O6-methylguanine-DNA-methyltranferase in a Series of 100 Glioblastoma Patients. *Cancer* 118 (17): 4201-4211.

Reifenberger, Guido, Bettina Hentschel, Jörg Felsberg, Gabriele Schackert, Matthias Simon, Oliver Schnell, Manfred Westphal, Wolfgang Wick, Torsten Pietsch, Markus Loeffler et al. 2012. Predictive Impact of MGMT Promoter Methylation in Glioblastoma of the Elderly. *International Journal of Cancer* 131 (6): 1342-1350.

Romano, Andrea, L. F. Calabria, F. Tavanti, G. Minniti, M. C. Rossi-Espagnet, V. Coppola, S. Pugliese, D. Guida, G. Francione, C. Colonnese et al. 2013. Apparent Diffusion Coefficient Obtained By Magnetic Resonance Imaging as a Prognostic Marker in Glioblastomas: Correlation with MGMT Promoter Methylation Status. *European Radiology* 23 (2): 513-520.

Seymour, L., S. P. Ivy, D. Sargent, D. Spriggs, L. Baker, L. Rubinstein, M. J. Ratain, M. Le Blanc, D. Stewart, J. Crowley, S. Groshen, J. S. Humphrey, P. West and D. Berry. 2010. The Design of Phase II Clinical Trials Testing Cancer Therapeutics: Consensus Recommendations from the Clinical Trial Design Task Force of the National Cancer Institute Investigational Drug Steering Committee. *Clinical Cancer Research* 16: 1764-1769.

Shah, Nameeta, Biaoyang Lin, Zita Sibenaller, Timothy Ryken, Hwahyung Lee, Jae-Geun Yoon, Steven Rostad and Greg Foltz. 2011. Comprehensive Analysis of MGMT Promoter Methylation: Correlation with MGMT Expression and Clinical Response in GBM. *PloS One*, 6 (1): e16146.

Simon, Richard. 2012. Clinical Trials for Predictive Medicine. *Statistics in Medicine* 31 (25): 3031-3040.

Simon, Richard M., Soonmyung Paik and Daniel F. Hayes. 2009. Use of Archived Specimens in Evaluation of Prognostic and Predictive Biomarkers. *Journal of the National Cancer Institute* 101 (21): 1446-1452.

Stupp, Roger, Warren P. Mason, Martin J. Van Den Bent, Michael Weller, Barbara Fisher, Martin J. B. Taphoorn, Karl Belanger, Alba A. Brandes, Christine Marosi, Ulrich Bogdahn et al. 2005. Radiotherapy Plus Concomitant and Adjuvant Temozolomide for Glioblastoma. *New England Journal of Medicine* 352 (10): 987-996.

Sunwoo, Leonard, Seung Hong Choi, Chul-Kee Park, Jin Wook Kim, Kyung Sik Yi, Woong Jae Lee, Tae Jin Yoon, Sang Woo Song, Ja Eun Kim, Ji Young Kim et al. 2013. Correlation of Apparent Diffusion Coefficient Values Measured By Diffusion MRI and MGMT Promoter Methylation Semiquantitatively Analyzed with MS-MLPA in Patients with Glioblastoma Multiforme. *Journal of Magnetic Resonance Imaging* 37 (2): 351-358.

Thon, Niklas, Sabina Eigenbrod, Eva M. Grasbon-Frodl, Juergen Lutz, Simone Kreth, Gabriele Popperl, Claus Belka, Hans A. Kretzschmar, Joerg-Christian Tonn and Friedrich W. Kreth. 2011. Predominant Influence of MGMT Methylation in Non-resectable Glioblastoma After Radiotherapy Plus temozolomide. *Journal of Neurology, Neurosurgery & Psychiatry* 82: 441-446.

Tunis, Sean R., Daniel B. Stryer and Carolyn 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.

Watanabe, Reiko, Yoko Nakasu, Hiroshi Tashiro, Koichi Mitsuya, Ichiro Ito, Satoshi Nakasu and Takashi Nakajima. 2011. O6-methylguanine DNA Methyltransferase Expression in Tumor Cells Predicts Outcome of Radiotherapy Plus Concomitant and Adjuvant Temozolomide Therapy in Patients with Primary Glioblastoma. *Brain Tumor Pathology* 28 (2): 127-135.

Weiler, Markus, Christian Hartmann, Dorothee Wiewrodt, Ulrich Herrlinger, Thierry Gorlia, Oliver Bähr, Richard Meyermann, Michael Bamberg, Marcos Tatagiba, Andreas von Deimling et al. 2010. Chemoradiotherapy of Newly Diagnosed Glioblastoma with Intensified Temozolomide. *International Journal of Radiation Oncology*Biology*Physics* 77 (3): 670-676.

Weller, Michael, Jörg Felsberg, Christian Hartmann, Hilmar Berger, Joachim P Steinbach, Johannes Schramm, Manfred Westphal, Gabriele Schackert, Matthias Simon, Jörg C. Tonn et al. 2009. Molecular Predictors of Progression-free and Overall Survival in Patients with Newly Diagnosed Glioblastoma: A Prospective Translational Study of the German Glioma Network. *Journal of Clinical Oncology* 27 (34): 5743-5750.

Wick, Wolfgang, Michael Weller, Martin van den Bent, Marc Sanson, Markus Weiler, Andreas von Deimling, Christoph Plass, Monika Hegi, Michael Platten and Guido Reifenberger. 2014. MGMT Testing—The Challenges for Biomarker-based Glioma Treatment. *Nature Reviews Neurology* 10: 372-385.

Worrall, John. 2010. Evidence: Philosophy of Science Meets Medicine. *Journal of Evaluation in Clinical Practice* 16: 356-362.

Spencer Phillips Hey is a postdoctoral fellow in the Studies of Translation, Ethics, and Medicine (STREAM) Research Group at McGill University. He received his doctorate in philosophy in 2011 from the University of Western Ontario. His work on the ethics and methodology of clinical trials has appeared in a variety of venues, both scientific and philosophical, including *Science Translational Medicine*, *Neurology*, *Philosophy of Science*, and the *Journal of Medical Ethics*.

Address: STREAM Research Group, McGill University, 3647 Peel St., Montréal QC H3A1X1, Canada. E-mail: spencer.hey@mcgill.ca