Perils of Evidence-Based Medicine

Ben A. Williams

ABSTRACT Evidenced-based medicine views random-assignment clinical trials as the gold standard of evidence. Because patient populations are heterogeneous, large numbers of patients must be studied in order to achieve statistically significant results, but the means or medians of these large samples have weak predictive validity for individual patients. Further, the logic of random-assignment clinical trials allows only the inference that some subset of patients benefits from the treatment. Post-hoc analysis is therefore necessary to identify those patients. Otherwise, many patients may receive treatments that are useless and potentially harmful.

Medicine has avidly embraced evidence-based medicine: best medical practices should be determined not by the experience of the individual clinicians, nor by the accumulated wisdom of clinicians more generally, but by the results of well-controlled clinical trials. Yet despite its increasing dominance in medical education, evidence-based medicine (EBM) has received considerable criticism, especially from philosophers of science who have questioned the central tenet that randomized trials should be the primary basis for establishing clinical practice (e.g., Bluhm 2005; Borgerson 2009; Goldenberg 2009; Grossman and Mackenzie 2005). Because this claim about randomized trials is essentially pragmatic in nature, it must be justified by a demonstration that better clinical outcomes result from randomized trials than from other types of clinical evidence. In fact, however, no such demonstration has been provided.

Whereas criticisms by philosophers of science have focused on the justifica-

Department of Psychology, University of California–San Diego, La Jolla, CA 92093-0109.
E-mail: bawilliams@ucsd.edu.

Perspectives in Biology and Medicine, volume 53, number 1 (winter 2010):106–20

tion of EBM's claim that some types of evidence are better than others, the present discussion focuses upon the actual implementation of EBM and various adverse consequences of that implementation. While EBM does encourage the judicious use of clinical evidence other than that from randomized trials, its emphasis on the importance of trials has fostered an often-uncritical acceptance of their evidence without an appreciation of their limitations. The result has been mistaken conclusions that have the potential for substantial harm.

THE GOLD STANDARD OF EVIDENCE

EBM's highest standard of evidence comes from large clinical trials (phase III) that randomly assign patients to the treatment condition or to a comparison control group. While new medical treatments may seem to show substantial benefits in nonrandomized phase II trials, in terms of the percentage of patients who respond to the treatment, this typically is not considered adequate evidence to make such treatments generally available. Randomized trials are required, because only with random assignment can the outcomes of the patients receiving the treatment be ascribed definitively to the treatment itself and not to unrepresentative features of the patient population that might give a favorable clinical outcome, regardless of the effects of the treatment per se.

The great majority of clinical trials adopt "null hypothesis testing" as the method of statistical evaluation. The null hypothesis posits that patients receiving the treatment have the same outcomes as those in the control group; this hypothesis must be disproved before the treatment can be considered effective. The starting assumption of null hypothesis testing is that patients participating in a phase III clinical trial represent the universe of patients within a given diagnostic category who might eventually receive the treatment. The standard criterion of plausibility is that the chance probability of the treatment group outcomes coming from the distribution of control group outcomes must be less than 5% ($p < .05$). Any outcome with a chance probability greater than .05 does not allow the null hypothesis to be rejected, and the result is that the clinical trial is considered a failure and the treatment is not validated for use. Conversely, if the probability of the treatment effect being due to random variability is less than .05, the treatment passes muster.

This model is so widely adopted that its conceptual foundations are seldom questioned. But while many scientific disciplines continue to use null hypothesis testing as their standard of evaluation, its validity has long been criticized (see, for example, Bakan 1966; Rozeboom 1960). Recently, alternatives have been proposed that address these longstanding issues and provide a methodology that better addresses the actual decisions underlying scientific inference (Killeen 2005).

The Information Content of Randomized Phase III Clinical Trials

A patient's primary concern when consulting a clinician is the benefit the clinician can provide. The details of that benefit are important, both in terms of the likely increase in survival, or clinical improvement, and whether that increase outweighs any loss in quality of life from the effects of the treatment. The fact that a phase III clinical trial allows the null hypothesis to be rejected provides little information for addressing this primary concern of patients, other than the statement that patients who received the treatment in the clinical trial had better outcomes, on average, than those who did not.

The null hypothesis model assumes two kinds of effects on clinical outcomes: those due to individual differences among patients, and those from the treatment. The task of evaluation is in essence a signal-to-noise problem, in which the study tries to pick out the signal value of the treatment from the statistical noise created by the individual differences between patients. If patients included in the disease category under study vary widely with respect to the many different characteristics that cause variable clinical outcomes (such as patient age), detecting the signal value of the treatment becomes substantially more difficult. And such heterogeneity may have profound effects on the power of the statistical assessment, with some estimates suggesting its power may be reduced by as much as 70% (Li et al. 2002).

A more fundamental problem with randomized clinical trials is that any conclusions derived from null hypothesis testing assumes that patients participating in the clinical trial are similar with respect to the effect of the treatment. In reality, however, a large part of the variability in clinical outcomes is due to the treatment effects varying with the characteristics of the individual patients. In other words, only some subsets of the patients in the clinical trial may benefit from the treatment. Especially since the revolution in genetic knowledge, this category of variance in treatment outcomes has assumed increasing importance. Even though the variance due to the interaction between treatment and individual difference variables is often large, this variability is generally regarded as statistical noise, again resulting in the main effect of the treatment being obscured.

Especially problematic is that a statistically significant outcome does not identify *which* patients will benefit from the treatment. Noted philosopher of science Nancy Cartwright (2007) argues that the logic of randomized trials permits only the conclusion that some subset of the patients in the clinical trial benefited from the treatment, which implies that post-hoc analysis of the results of individual patients is required before clinical recommendations can be made for individual patients. While post-hoc analysis is often performed, it is primarily used to guide future clinical trials, rather than to determine whether the treatment under study is beneficial or harmful.

Given the above considerations, what information do phase III clinical trials provide that would aid the patient's decision about treatment? If the clinical trial

fails to meet the criterion of statistical significance, it does not mean that the treatment is ineffective, merely that a significant effect of the treatment has not been demonstrated. Significant effects fail to occur for a number of reasons, including highly heterogeneous patient populations and poor execution of the trial protocol. Moreover, as will be exemplified later, acceptance of the null hypothesis routinely occurs in medical clinical research, even though one of the first lessons taught to students in introductory statistics is that such acceptance is unwarranted.

Even if a randomized clinical trial does achieve a statistically significant effect, one that is sufficient to obtain Food and Drug Administration (FDA) approval for the drug or treatment to be marketed, such a result may have minimal predictive utility at the level of the individual patient. Due to the statistical noise, large numbers of patients typically must be included in the clinical trial. But given that a statistically significant effect allows only the conclusion that some subset of patients benefits from the treatment, individual patients have little basis for determining whether they are among that subset. It is only by comparing their own characteristics to those of patients in the clinical trial that the new patients have a meaningful basis for inferring that the new treatment may be worthwhile for them.

Another, separate limitation on the information content of randomized clinical trials is that they rarely involve patients who are a random sample of those who will potentially receive the treatment. Recruitment of patients into clinical trials is often difficult, and a large percentage of patients who do participate are those most desperate due to previous treatment failures. Thus, clinicians must consider not only the outcomes of the clinical trials but also whether their own patients are similar to those in the trial. Although lip service is paid to the limitation on the generalizability of the clinical trial results, in actual clinical practice this limitation is often ignored.

It is useful to consider the generalizability of results from phase III trials in relation to the supposed limitations of phase II trials, which do not involve randomization and control groups. The essential deficiency of phase II trials is that their outcomes may result from the idiosyncratic characteristics of the patient populations being studied, and not necessarily from the treatment itself; thus, the results may not be applicable to other patients. Phase III trials avoid this problem through randomization, so that any differential outcome for the treatment vs. control groups can be ascribed only to whether the patients received the treatment. But due to the uncertainty about whether the patients participating in the phase III trial are in fact representative of the patients who will potentially receive the treatment, *and* the possibility that individual patient characteristics may still strongly affect treatment outcome, the supposed advantages of phase III over phase II trials are substantially diminished. Neither type of trial eliminates the concern that its results will fail to generalize beyond the patients participating in the trial.

Rigid Adherence to "Statistical Significance" Distorts Treatment Evaluation

The generally accepted convention for statistical significance is that the outcome of a clinical trial must have a chance probability level less than .05. This means that the outcome of the clinical trial, in terms of differences in central tendency such as the median or mean, has less than one chance in 20 of being due to random variability. Such conventions are entirely arbitrary, and there is no reason why the same standard should be used for all situations. It is important to keep in mind that the .05 threshold was borrowed from other, nonclinical, sciences, in which the bias is to ensure that scientific observations are real empirical facts. But such a conservative bias is inappropriate in many clinical settings, especially for diseases without effective treatment. Why, for example, should a probability level of .15 not be sufficient for approval of a new treatment, given that this means there is only a 15% chance that the difference in outcome was due to random error and not a real effect? If the patient has essentially no other options, and meaningful evidence from nonrandomized phase II trials suggests that the treatment under investigation offers a clinical benefit, denying access to the treatment based on its failure to achieve the arbitrary .05 level of statistical significance seems foolish at best, and arguably inhumane.

The more fundamental problem is that demonstrating a *statistically* significant benefit of a new treatment may provide only weak evidence that the benefit is *clinically* significant. An increasing proportion of clinical trials use very large numbers of patients in order to achieve the criterion of statistical significance. In fact, some have argued that it is unethical to conduct small clinical trials, because they are unlikely to achieve statistical significance: patients' participation in such trials has been characterized as a futile enterprise unlikely to yield any useful information (Halpern et al. 2002). Such criticisms reflect the hegemony of the concern for statistical significance.

Meeting a criterion of statistical significance provides no solid foundation for believing that a treatment is clinically useful. When a clinical trial establishes that a new treatment produces a reliable increase in median survival of a few months, this provides a weak basis for any given patient's use of that treatment, especially if accompanied by significant side effects. The clinically relevant result of the clinical trial is instead the *effect size*, namely, how much of the variability in clinical outcomes is due to having received the treatment or not. For many areas of medicine, especially oncology, effect sizes typically are dismally small. Indeed, the overlap in clinical outcomes for patients receiving the treatment or not is often so great that patients have minimal information about whether the treatment will provide a benefit for them as individuals, and especially whether the benefit from the treatment outweighs the loss in quality of life engendered by the treatment's toxicity.

The critical fact, too often ignored, is that increases in the number of participants in a trial in no way alters the effect size, and it is the effect size, not the

probability level of the statistical test, that is the clinically relevant information. Contrary to the view just cited that "small clinical trials are unethical," the opposite inference follows from the actual nature of clinical trial results. Small clinical trials should be favored, because only then will it be possible to infer with confidence that a statistically significant result is meaningfully predictive of the treatment being effective at the level of the individual patient.

The information provided by a statistically significant effect in a randomized trial is that patients in the treatment group have reliably different outcomes, in terms of measures of the mean or median, than patients in the control group. That is, if comparable samples of patients were drawn from the original population of patients, the superiority of the treatment over the control condition would very likely be replicated (but not necessarily the size of the difference). But reliability at the level of samples of patients is not the clinically relevant information. If large samples of patients are necessary to produce a statistically reliable effect, the corollary is that the mean or median result of the sample has weak predictive validity for the individual patient. In other words, if one were trying to predict the outcome of a treatment for a specific individual patient, a prediction based on the results of a large phase III clinical trial often would be only marginally better than simply guessing.

It is important to appreciate that null hypothesis testing is not the only possible statistical method of clinical evaluation. Dominance statistics provide one alternative, as they provide the probability that a randomly drawn patient from the treatment condition will exceed in outcome a randomly drawn patient from the control condition (Bamber 1975). Such information is clinically more valuable than knowing that a large sample of patients in the treatment condition has met the .05 criterion of having a better outcome than a large sample of control patients.

CONSEQUENCES OF FAILING TO RECOGNIZE THE PERILS

The foregoing issues are not mere abstractions, but rather are found in many everyday medical decisions. The examples to be considered involve several different clinical issues.

The Cost of Accepting the Null Hypothesis

Until only recently, a controversial issue in the treatment of brain cancer was whether chemotherapy provided any benefit. The outcomes of numerous clinical trials were highly variable, with the phase II trials frequently indicating a benefit while randomized phase III trials did not. The Medical Research Council Brain Tumour Working Party (2001) conducted a null effect phase III trail in Canada and the United Kingdom, in which 674 patients with high-grade gliomas (including both glioblastomas and anaplastic astrocytomas) were ran-

domly assigned to receive either radiation as the sole treatment or radiation in combination with PCV chemotherapy (a combination of procarbazine, lomustine, and vincristine). The results were a median survival of 9.5 months for radiation alone and 10 months for radiation plus PCV, a difference that did not approach statistical significance.

If these results had been generally accepted, they would have resulted in high-grade glioma patients being offered nothing beyond radiation therapy as treatment. However, several features of the results deviated so significantly from previous phase II clinical trials that various critics questioned their validity. Not only did the trial show no benefit of chemotherapy for glioblastoma tumors (not a surprising result), but also none for anaplastic astrocytomas, which had been regarded as more responsive to chemotherapy. Moreover, the survival time for patients with anaplastic astrocytomas was only 13 months, much shorter than typically had been observed in previous clinical trials. For example, a different large clinical trial conducted to test the effect of a radiation sensitizer (which failed to provide a benefit) reported a median survival time of 45 months (Prados et al. 2004).

Given the disparity between the Medical Research Council trial and previous anaplastic astrocytoma results, as well as various other discordant features of the results (such as the failure to find any effects of well-established prognostic variables, such as age and Karnofsky score), the clinical issue was whether the null result in the randomized phase III trial should override the seemingly much more positive results in the previous phase II trials. This particular large phase III trial had little influence because of various shortcomings in the trial execution (Chamberlain and Jaeckle 2001), but while possible reasons for the null effect reported in this trial could be identified, it is important to recognize that all clinical trials with null effects face a similar problem of interpretation.

A related example comes from two separate clinical trials involving oligodendroglioma tumors (Cairncross et al. 2006; Van den Bent et al. 2006). The design of both trials was essentially similar, a comparison of radiation (RT) alone versus RT plus PCV chemotherapy. The conclusion of both studies was that there was no effect of the addition of PCV in terms of overall survival: median survival in one study was 40.3 months for the RT plus PCV versus 30.6 months for RT alone ($p = .23$); median survival for the other study was 58.8 months for RT plus PCV versus 56.4 months for RT alone ($p = .26$). However, both studies also reported a statistically significant effect of PCV on the measure of progression-free survival, 23 versus 13.1 months in one case, 31.2 versus 20.4 months for study two.

Given that progression-free survival and overall survival are usually highly correlated, what then accounts for the differences between the two outcome measures? In both clinical trials patients whose tumor progressed then received additional chemotherapy (with a different chemotherapy agent if they initially

received PCV, or PCV chemotherapy for the first time if they initially received only radiation). Such "salvage therapy" has highly variable results across patients, which may vary with a patient's prior history of chemotherapy. Given this additional source of variance (part of which was confounded with the actual treatment variable under study), it should not be surprising that the significant differences obtained with the most uncontaminated measure of treatment effectiveness were overshadowed. This problem is now common in oncology because of increased availability of second-line treatments: patients in clinical trials almost always receive some type of salvage therapy after the initial treatment under investigation has failed. The resulting increase in statistical noise can easily obscure the main effect of the primary treatment under study, causing false negatives to be substantially more likely.

Yet a third example of the issues raised by statistical outcomes that do not meet the standard .05 level of significance involves the decision process of the FDA. The current standard of care for glioblastoma brain tumors is temozolomide (trade name Temodar). After several phase II trials suggesting that it provided better results that traditional chemotherapy protocols such as BCNU (Carmustine) or PCV, patients whose tumors had progressed after prior treatment with either BCNU or PCV were randomized to receive either temozolomide or procarbazine (Yung et al. 2000). Several different measures of clinical outcome were reported: (1) the percentage of patients who had no tumor progression for at least six months after initiation of treatment (21% for temozolomide versus 9% for procarbazine); (2) the median time between the start of treatment and tumor progression (2.9 months versus 1.9 months); and (3) the median survival time after treatment initiation (7.3 months versus 5.8 months). The first two differences were statistically significant using the $p < .05$ criterion, but the difference in overall survival failed to reach the .05 criterion.

One likely reason for why the difference in overall survival time did not reach conventional significance levels was the statistical noise due to individual difference variables (such as age and Karnofsky score), which produce variation in survival time regardless of treatment. When the patients were partitioned into high versus low categories for each of the major prognostic variables, in every case there was a difference in favor of the temozolomide condition. However, despite the overall pattern of results, the FDA refused to approve temozolomide as a treatment for glioblastomas. Such a decision can only be viewed as a slavish endorsement of that one statistical criterion as the ultimate arbiter of clinical effectiveness, without regard for the alternative treatment possibilities and the overall pattern of evidence. The cost of this decision was delayed access to a new drug that subsequently has been shown to improve clinical outcomes for brain cancer patients.

The Critical Importance of Individual Differences

A major motivation for modern medicine's endorsement of randomized phase III clinical trials as the gold standard of evidence is the recurrence of studies in which phase II trials indicate a major benefit from a new treatment agent, but follow-up randomized phase III trials fail to produce a statistically significant effect. While this pattern usually is interpreted as showing that phase II clinical trials are contaminated by unrepresentative patient populations, an alternative explanation is that the patient populations participating in phase III clinical trials are so heterogeneous with respect to major prognostic variables that the main effect of the treatment becomes difficult to detect.

To illustrate the magnitude of the variability, consider the landmark clinical trial that resulted in temozolomide being FDA approved for the treatment of glioblastoma multiforme brain cancer. In this large multi-center European trial (Stupp et al. 2005), 573 patients were randomized to receive only radiation or radiation in combination with temozolomide, first at daily low doses during radiation, and then on a schedule of days one to five of every month. Median survival was 14.6 months for RT plus temozolomide, versus 12.1 months for RT alone, a difference that attained statistical significance. More impressive was the difference in two-year survival rate: 26 % for the combination, but only 10% for RT alone. Consider the information content of this trial from the perspective of the individual patient. Because the difference in median survival time was only 2.5 months, and the fact that over 500 patients were necessary to demonstrate a statistically significant effect, the difference in median survival time provided weak evidence for the individual patient to decide to receive the additional treatment along with its accompanying toxicity.

Far more informative for the decision of the individual patient were the results of a follow-up analysis of how clinical outcomes were affected by a specific genetic marker, whether the MGMT DNA-repair enzyme was silenced by methylation of its promoter (Hegi et al. 2005). The rationale for this analysis was that tumor cell damage caused by the chemotherapy agent could not be repaired if the MGMT gene were inactivated. Median survival for patients with the inactive MGMT gene was 22 months for patients receiving chemotherapy versus 15 months for RT alone. The corresponding results for two-year survival were 46% and 22.7%. In contrast, for patients with an activated MGMT gene, the median survivals were 12.7 versus 11.8 months for the combination versus radiation-alone conditions, whereas their corresponding two-year survivals were 14% versus 2%.

The major lesson to be learned from this retrospective analysis is that a large percentage of the variance in the treatment outcome measures can be accounted for by a single individual difference variable, both in terms of that variable having independent prognostic status and in terms of it predicting treatment efficacy. Because this individual difference factor was treated as statistical noise in the randomized trial, it was necessary to use an extremely large number of pa-

tients, as a smaller N likely would have been problematic with respect to attaining the critical $p < .05$ criterion. Moreover, the main effect of the treatment protocol was small relative to the impact of the individual difference variable, which was discovered by the post-hoc analysis of the effect of the MGMT gene status. That information is critical for patients with the active gene, because it implies that temozolomide is unlikely to provide them any benefit, and that they are better served by seeking a different treatment protocol.

A second example concerns one of the most controversial issues in oncology—whether cancer patients should use nutritional supplements, specifically those that are antioxidants. Conventional oncologists typically recommend against such use, while advocates of alternative/complementary medicine have generally favored it. A major development in adjudicating this issue was a large clinical trial involving patients with head-and neck cancer, who were treated with conventional radiation therapy (Bairati et al. 2006). Patients (N =540) were randomized to receive either radiation and placebo, or radiation and beta-carotene and vitamin E. By the end of an eight-year follow-up, 77 patients receiving placebo had died, while 102 patients receiving the antioxidant supplements had died, a difference that was statistically significant. This finding received widespread publicity and was taken as strong vindication for the view that supplements were harmful for cancer patients receiving treatment, although the clinical trial also showed that the antioxidants produced substantial alleviation of the debilitating effects of radiation (such as mucositis, xerostomia, and weight loss). However, the seemingly clear results of the clinical trial were undercut by a post-hoc analysis showing that all of the increased mortality due to the antioxidant supplementation was confined to the subgroup of patients who continued to smoke while receiving radiation (Meyer et al. 2008). Those who did not smoke, including those who previously had smoked, had no increase in death rate. Had the post-hoc analysis not been performed, the result would have been that cancer patients would be strongly advised to refrain from using antioxidant supplements and thus would needlessly endure the very debilitating side effects of the radiation treatment. Moreover, smokers who continued smoking during radiation treatment would not have been alerted to the extreme importance of their not using antioxidants during that period.

The issues discussed above arise repeatedly in many medical treatments. A final example from a much more common medical condition, atrial fibrillation, provides an illustration. Atrial fibrillation is the most frequent problem presented to cardiologists, with estimates of its rate of occurrence sometimes as high as 10% of the population. While it occurs disproportionally in the elderly, it also occurs among those relatively young. Often it occurs concomitant with other heart problems, such as valvular dysfunction and prior heart disease, but often also for people who seem to have no other cardiac pathology. Thus, there is substantial variation in the medical histories of those with the diagnosis.

For many years the most common treatment for atrial fibrillation was quinidine. Beginning in the late 1980s, this quickly changed because several clinical trials indicated that quinidine usage was associated with an increased risk of sudden death, most frequently from the development of torsades des pointes. At that point, the consensus among cardiologists was that quinidine did more harm than good, a conclusion that was generalized to other anti-arrhythmic drugs.

However, as noted above, patients with atrial fibrillation have diverse clinical histories, including various other cardiac problems. In one review of eight different clinical trials with quinidine, for example, all trials had at least 50% of patients with some form of structural heart disease, and in the great majority the incidence was 65 to 90% (Lafuente-Lafuente 2006). Most informative is a randomized clinical trial that separated patients according to history of congestive heart failure (Flaker et al. 1992). For those with such history, the rate of cardiac death was 4.7 times higher for those receiving quinidine versus those receiving the placebo. For those without a history of congestive heart failure, quinidine decreased the incidence of cardiac death by 30%. Thus, depending on the patient population, quinidine is either a bad drug or a drug with considerable effectiveness. But despite its effectiveness for patients without structural heart disease (the majority of patients who potentially might use it), quinidine is now virtually banned as a drug.

Historical Controls versus the Gold Standard

Whatever their limitations, phase III randomized clinical trials do have one property that proponents of evidenced-based medicine regard as essential: they provide a basis for establishing that a treatment agent does provide a real benefit for at least some of the patients receiving the treatment. Thus, phase III trials provide proof that a given treatment is more than mere snake oil. But is it possible to pass the snake oil test without randomized trials?

In terms of actual medical practice the answer to this question is apparently yes. A large portion of current medical practice involves off-label drug use, which has rarely been certified by randomized trials. This widespread practice raises questions about whether off-label drug use is illegitimate from the perspective of EBM and, if not, what its implications are for the status of phase III randomized trials as the highest standard of evidence.

While the diversity of off-label drug use precludes any well-defined rules of how such use is determined, the great majority results from nonrandomized phase II trials, in which clinical outcomes have been compared to some form of historical controls. An example of such use again comes from clinical trials for glioblastoma. After years of failed phase III clinical trials, a historical record of six-month progression-free survival (PFS-6) was compiled from previous clinical trials involving patients with recurring tumors, for whom the prognosis is

very poor (Wong et al. 1999). The result was a dismal PFS-6 value of 15%. Results from subsequent phase II trials have been compared to this benchmark to identify which treatment agents have potential for further development. The result has been that a variety of new treatment agents, primarily those already FDA-approved for other purposes, have become widely used, greatly expanding the treatment options for a patient population in dire need of new treatments. In the most notable case, the anti-angiogenic drug avastin has been approved by the FDA for the treatment of glioblastoma, based not on a randomized phase III trial, but on the results of multiple phase II trials in which the PFS-6 values were markedly superior to the historical norm.

The use of PFS-6 values is a relatively crude use of historical controls, especially given its lack of differentiation with respect to individual differences between patients. A more sophisticated use of historical controls has been enabled by the identification of six categories of patients in the corpus of previous clinical trials (Curran et al. 1993; Scott et al. 1998), defined by combinations of eight different prognostic variables that have substantially different clinical outcomes. Only categories III through VI are commonly used, because they correspond to patients with glioblastomas. Median survival times range from 17.9 months for category III to only 4.6 months for category VI. Patients in each new phase II clinical trial can thus be compared to their appropriate category matched to their prognostic features.

The value of this approach is shown by an analysis of the effects of brachytherapy, involving the implantation of radiation seeds in the tumor area inside the brain (Videtic et al. 1999). Numerous phase II trials have reported an apparent survival gain of six to 12 months, but a large phase III trial failed to demonstrate a significant benefit (Selker et al. 2002). The benefits reported in phase II clinical trials have accordingly been regarded as due to selection bias. However, when patients receiving brachytherapy were partitioned according to prognostic variables and compared to historical controls within the same category, brachytherapy produced longer median survival times and greater two-year survival rates within all four of the prognostic categories. The fact that all patient groups receiving brachytherapy fared better than their comparable historical controls severely questions any interpretation of the benefits of brachytherapy in terms of bias due to subject selection. Should this outcome or the null result from the large randomized trial guide clinical practice?

The advantages of basing drug approval decisions on phase II trials, in conjunction with historical controls, are obvious: a reduced time to drug approval, and a greatly reduced expense of treatment development. The disadvantages are more contentious. Critics argue that historical controls are not necessarily comparable to the patients in any given phase II trial, and that attempts to use them in the past have failed (although the criterion of failure is unclear). Moreover, due to general improvements in medical outcomes, historical controls are moving targets, so that any improvement in clinical outcome relative to outdated his-

torical controls will overestimate treatment benefits and potentially endorse treatments as effective that in reality may not be better than a placebo.

While the concerns about the adequacy of historical controls are significant, it is important to recognize that previous uses of this approach had much less extensive databases and often were based on matching individual control subjects to the experimental subjects on a limited number of prognostic variables. The approach advocated here uses categories of patients as the historical controls, with the categories defined by many more prognostic variables, including various genetic markers. Such usage would require frequent updating of the historical database, but this would be a far less expensive undertaking that the cost of large randomized clinical trials.

Certainly all of the relevant variables that determine clinical outcome will never be known, so any sampling bias in a given phase II trial may potentially produce a positive outcome that is not representative of the general patient population. But this problem must be weighed against the problem inherent in phase III trials of overgeneralizing the benefits of the treatment. Because patients included in any diagnostic category are not homogeneous, the results of the trials may be applicable only to specific subgroups of patients, resulting in many patients, perhaps a majority, receiving useless treatments that are possibly harmful. The pivotal issue is the amount of variance accounted for by using individual patient profiles based on the historical record versus the amount of variance in randomized trials that is due to patient heterogeneity. The use of the historical record has the advantage that its predictive power can be incrementally improved, while the problems of heterogeneous patient populations in randomized trials are more intractable, unless such trials include much more post-hoc (or possibly preplanned) subgroup analyses. However, such analyses are believed by many to be invalid because they subvert the benefits of randomization, unless each subgroup can itself be randomly assigned to the treatment versus control conditions.

Individualized treatment protocols have received increasing discussion in the medical literature, and they will receive even more as further progress is made in identifying the genetic and epigenetic markers that predict differential outcomes. The use of historical controls that have been differentiated according to multiple prognostic factors is a first step toward developing such individualized treatments. It is much more difficult to imagine how large randomized trials will aid this enterprise.

As noted in the introduction, the claim that a randomized clinical trial is the best strategy for advancing clinical medicine is an empirical issue, one that should be adjudicated by the actual effects of adopting that strategy. The examples presented above provide evidence that the strategy may produce various undesirable consequences that must be weighed against the major benefit of EBM, which is to provide unequivocal evidence that a given treatment actually

does benefit at least some patients. The outcome of that evaluation is by no means a foregone conclusion.

REFERENCES

Bairati, I., et al. 2006. Antioxidant vitamins supplementation and mortality: A randomized trial in head and neck cancer patients. *Int J Cancer* 119:2221–24.

Bakan, D. 1966. The test of significance in psychological research. *Psychol Bull* 66:423–37.

Bamber, D. 1975. The area above the ordinal dominance graph and area below the receiver operating characteristic graph. *J Math Psychol* 12:387–415.

Bluhm, R. 2005. From hierarchy to network: A richer view of evidence for evidence-based medicine. *Perspect Biol Med* 48(4):535–47.

Borgerson, K. 2009. Valuing evidence: Bias and the evidence hierarchy of evidence-based medicine. *Perspect Biol Med* 52(2):218–33.

Cairncross, G., et al. 2006. Phase III trial of chemotherapy plus radiotherapy compared with radiotherapy alone for pure and mixed anaplastic oligodendroglioma: Intergroup Radiation Therapy Oncology Group Trial 9402. *J Clin Oncol* 24(18):2707–14.

Cartwright, N. 2007. Are RCTs the gold standard? *Biosocieties* 2(2):11–20.

Chamberlain, M. D., and K. A. Jaeckle. 2001. Medical Research Council adjuvant trial in high-grade gliomas. *J Clin Oncol* 19(19):3997–98.

Curran, W. J., et al. 1993. Recursive partitioning analysis of prognostic factors in three Radiation Therapy Oncology Group malignant glioma trials. *J Natl Canc Inst* 85:704–10.

Flaker, G. C., et al. 1992. Antiarrhythmic drug therapy and cardiac mortality in atrial fibrillation: The stroke prevention in atrial fibrillation investigators. *J Am Coll Cardiol* 20(3):527–32.

Goldenberg, M. J. 2009. Iconoclast or creed? Objectivism, pragmatism, and the hierarchy of evidence. *Perspect Biol Med* 52(2):168–87.

Grossman, J., and F. J. Mackenzie. 2005. The randomized controlled trial: Gold standard or merely standard? *Perspect Biol Med* 48(4):516–34.

Halpern, S. D., J. H. Karlawish, and J. A. Berlin. 2002. The continuing unethical conduct of underpowered clinical trials. *JAMA* 288:358–62.

Hegi, M. E., et al. 2005. MGMT gene silencing and benefit from temozolomide in glioblastoma. *N Engl J Med* 352(10):997–1003.

Killeen, P. R. 2005. An alternative to null hypothesis significance tests. *Psychol Sci* 16:345–53.

Lafuente-Lafuente, C., et al. 2006. Antiarrhythmic drugs for maintaining sinus rhythm after cardioversion of atrial fibrillation: A systematic review of randomized controlled trials. *Arch Intern Med* 166(7):719–28.

Li, Y., et al. 2002. The use of frailty hazard models for unrecognized heterogeneity that interacts with treatment: Considerations of efficiency and power. *Biometrics* 58:232–36.

Medical Research Council Brain Tumour Working Party. 2001. Randomized trial of procarbazine, lomustine, and vincristine in the adjuvant treatment of high-grade astrocytoma: A Medical Research Council trial. *J Clin Oncol* 19(2):509–18.

Meyer, F., et al. 2008. Interaction between antioxidant vitamin supplementation and cigarette smoking during radiation therapy in relation to long-term effects on recurrence and mortality: A randomized trial among head and neck cancer patients. *Int. J. Cancer* 122:1679–83.

Prados, M. D., et al. 2004. Phase III randomized study of radiotherapy plus procarbazine, lomustine, and vincristine with or without BUdR for treatment of anaplastic astrocytoma: Final report of RTOG 9404. *Int J Radiat Oncol Biol Phys* 58(4):1147–52.

Rozeboom, W. 1960. The fallacy of the null-hypothesis significance test. *Psychol Bull* 57: 416–28.

Scott, C. B., et al. 1998. Validation and predictive power of Radiation Therapy Oncology Group (RTOG) recursive partitioning analysis classes for malignant glioma patients: A report using RTOG 90-06. *Int J Radiat Oncol Biol Phys* 40(1):51–55.

Selker, R. G., et al. 2002. The Brain Tumor Cooperative Group NIH Trial 87-01: A randomized comparison of surgery, external radiotherapy, and carmustine versus surgery, interstitial radiotherapy boost, external radiation therapy, and carmustine. *Neurosurgery* 51:355–57.

Stupp, R., et al. 2005. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N Engl J Med* 352(10):987–96.

Van den Bent, J. J., et al. 2006. Adjuvant procarbazine, lomustine, and vincristine improve progression-free survival but not overall survival in newly diagnosed anaplastic oligodendrogliomas and oligoastrocytomas: A randomized European Organisation for Research and Treatment of Cancer Phase III trial. *J Clin Oncol* 24(18):2715–22.

Videtic, G.M. M., et al. 1999. Use of the RTOG recursive partitioning analysis to validate the benefit of iodine-125 implants in the primary treatment of malignant gliomas. *Intl J Rad Onc Biol Phys* 45(4):687–92.

Wong, E. T., et al. 1999. Outcomes and prognostic factors in recurrent glioma patients enrolled onto phase II clinical trials. *J Clin Oncol* 17(8):2572–78.

Yung, W. K., et al. 2000. A phase II study of temozolomide vs. procarbazine in patients with glioblastoma multiforme at first relapse. *Br J Cancer* 83(5):588–93.