Higher Field Strength for Proton MR Spectroscopy

MR imagers with magnetic fields (B_0) greater than 1.5 T are offered by all major manufacturers. Although 4-, 6-, 7-, and even 8-T, whole-body instruments are currently operational, the most common high-B_0 systems are the nearly 100 installed with 3-T magnets. The demand for bigger B_0 systems has been driven almost exclusively by functional MR imaging (1); however, their proliferation and Food and Drug Administration approval for most MR systems raises the question of whether the associated cost and technical complications benefit other applications such as MR spectroscopy. Indeed, several comparisons were already reported on the signal-to-noise ratio (SNR) and spectral resolution improvements realized from raising B_0 from the ubiquitous 1.5 T to 3, 4, and 7 T (2–5); however, common to all were that 1) few (<10) healthy subjects were examined; therefore, 2) none evaluated capability to systematically identify or differentiate abnormal states; 3) none observed the theoretical gains ($\propto B_0$) in either attribute; and 4) proton MR spectroscopy in the brain was used, which requires sequences of various echo times, fat, and water suppression.

In this issue of the *AJNR,* Kantarci et al tackle, for the first time, the first two above-referenced objectives. This was done with single-voxel proton MR spectroscopy in the posterior cingulate gyri in 20 subjects with mild cognitive impairment, a condition thought to precede Alzheimer disease; 20 with symptoms consistent with Alzheimer disease; and 41 age-matched control subjects. All 81 subjects underwent proton MR spectroscopy at 1.5 and 3 T. The metrics compared were ratios to creatine (Cr), a metabolite reflecting high-energy phosphate reserves in the cytosol of neurons and glial cells, of the neurotransmitters glutamine (Gln) and glutamate (Glu), and *myo*-inositol (MI), a marker of gliosis, obtained at short (30-millisecond) echo times; and *N*-acetylaspartate (NAA), a neuronal marker, and choline (Cho), a membrane turnover indicator, acquired at intermediate (135-millisecond) echo time. NAA/Cr and Cho/Cr, were also acquired and included in the analysis of the short–echo time experiment. The authors' goal was to ascertain which metabolic characteristics, field strengths, and echo times were most appropriate to differentiate the three subgroups.

Eighty-one quality-control acquisitions on a brain-metabolite phantom, the "best-case scenario" taken at both field strengths, provided an early indication of findings: the coefficients of variations (CV) of the metabolite ratios (to Cr) were lower and less variable at 1.5 T. It is no surprise, therefore, that this trend was repeated in the 41 elderly control subjects; only the Cho/Cr ratios were statistically different. Because the CV = SD/mean, assuming the SD of the measurements is approximately the same at each field strength, the finding that CV at 3 T is not half that at 1.5 T indicates that the theoretical signal intensity gains were not approached. In fact, with an echo time of 30 milliseconds (ie, a sequence providing minimal T2 loss), the SNR of Cr at 3 T was only 23% better than that at 1.5 T, whereas the line width more than doubled.

At 1.5 T, metabolite ratios of moderately cognitively impaired patients fell, as expected, between the elderly control subjects and those with Alzheimer disease, an additional indication of the intermediate nature of this condition. Specifically, MI/Cr and Cho/Cr acquired at echo times of 30 milliseconds progressively and significantly ($P < .05$) higher in control subjects as compared with ratios acquired in subjects with moderate cognitive impairment and Alzheimer disease. The NAA/Cr acquired at an echo time of 135 milliseconds showed progressively lower NAA/Cr and NAA/MI differentiated moderately cognitively impaired subjects from those with Alzheimer disease. By contrast, at 3 T, no metabolite ratios differentiated control subjects from those with mild cognitive impairment, or the latter subjects from those with Alzheimer disease. Although a trend of decreasing (Glu+Gln)/Cr was observed in moderately cognitively impaired subjects and those with Alzheimer disease, and was most pronounced in the latter group, the decrease did not reach statistical significance even in this large cohort. Consequently, Kantarci et al conclude that, in light of the current technology, 3-T proton MR spectroscopy offers no diagnostic performance advantage over the venerable 1.5-T field strength when applied to differentiating pathologic metabolism in the elderly.

Two technical causes were identified for this shortfall: 1) shimming (only first order *x, y, z* was used at either field strength) and 2) the effective contraction of T2s with field strength. Indeed, the posterior cingulate gyri were probed instead of the hippocampus or entorhinal cortex because of the poorer B_0 field homogeneity at the latter regions. Not discussed was the quantification by means of metabolite ratios rather than absolute metabolite concentrations. Although ratios benefit from cancellation of difficult to measure multiplicative factors, such as voxel partial CSF volume, instrumental gain, and interpersonal coil loading differences, their variation is the sum of individual components. Furthermore, ratios are also implicitly assumed to reflect the behavior of the numerator, because the denominator's level (frequently Cr) is presumed to be constant. This assumption was recently criticized by Weiner et al (6), who argued that absolute Cr level variations, especially as a function of age, may have also contributed to the lack of differential correlation between ratios and clinical status.

Because the theoretical SNR gains from bigger B_0s were already demonstrated for MR spectroscopy of other nuclei by the Minnesota group (7, 8), Kantarci et al assert that technical, not fundamental obstacles, impeded their achievements for proton MR spectroscopy. Therefore, it is left to academic hardware and technique developers to produce and to commercial manufacturers to bundle the solutions required to heed these requests from clinicians if big-B_0 imagers are to advance beyond fMR imaging applications.

ODED GONEN
Editorial Board

References

1. Thulborn KR. **Why neuroradiologists should consider very-high-field magnets for clinical applications of functional magnetic resonance imaging.** *Top Magn Reson Imag* 1999;10:1–2
2. Bartha R, Drost DJ, Menon RS, Williamson PC. **Comparison of the quantification precision of human short echo time (1)H spectroscopy at 1.5 and 4.0 Tesla.** *Magn Reson Med* 2000;44:185–192
3. Barker PB, Hearshen DO, Boska MD. **Single-voxel proton MRS of the human brain at 1.5 T and 3.0 T.** *Magn Reson Med* 2001;45:765–769
4. Gonen O, Gruber S, Li BSY, Mlynárik V, Moser E. **Multivoxel 3D proton spectroscopy in the brain at 1.5 versus 3 Tesla: a signal-to-noise and resolution comparison.** *AJNR Am J Neuroradiol* 2001;22:1727–1731
5. Vaughan JT, Garwood M, Collins CM, et al. **7T vs. 4T: RF power, homogeneity, and signal-to-noise comparison in head images.** *Magn Reson Med* 2001;46:24–30
6. O'Neill J, Schuff N, Marks WJ, Jr, et al. **Quantitative 1H magnetic resonance spectroscopy and MRI of Parkinson's disease** *Mov Disord* 2002;17:917–927
7. Zhu XH, Merkle H, Kwag JH, et al. **17O relaxation time and NMR sensitivity of cerebral water and their field dependence** *Magn Reson Med* 2001;45:543–549
8. Lei H, Zhu XH, Zhang XL, et al. **In vivo 31P magnetic resonance spectroscopy of human brain at 7 T: an initial experience** *Magn Reson Med* 2003;49:199–205

Glial Neoplasms without Elevated Choline-Creatine Ratios

The role of proton MR spectroscopy (MRS) in the differential diagnosis of a brain lesion often centers on cautiously assessing the choline (Cho) and total creatine (Cr) resonances and the choline-creatine ratio (Cho/Cr) in spectra from the lesion. In this issue of the *AJNR*, Londono et al and Saraf-Lavi et al underscore the pitfall of excluding neoplasm from the differential diagnosis when there is no elevation of Cho or Cho/Cr. The results from both groups of investigators indicate that some glial neoplasms, which present with variable mass effect and a lack of postcontrast enhancement on MR images, may not show elevation of Cho or Cho/Cr relative to values from normal-appearing brain, yet do show a marked increase in the 3.5–3.6 ppm resonance assigned to myo-Inositol (m-Ins) or glycine (Gly) or both. Evidence favoring m-Ins as the metabolite responsible for the increase may be gleaned from evaluation of both short TE and long TE spectra, and possibly from changes in the ~3.35 ppm scyllo-Inositol (s-Ins) resonance. Thus, other spectral measures should be sought before eliminating neoplasm from the differential diagnosis for a brain lesion with certain imaging features and without elevation of Cho or Cho/Cr.

For intracerebral lesions with variable mass effect and postcontrast enhancement, short TE (~ 20–30 ms) and longer TE (~ 135 ms) spectra showing *marked* elevation of Cho or Cho/Cr, and diminished *N*-acetyl (NA) or NA/Cr, favor a diagnosis of neoplasm over infection, inflammation, ischemia, or infarction. (Note: NA represents the combined resonances of *N*-acetyl aspartate [NAA] and *N*-acetyl aspartylglutamate [NAAG]. NAA and NAAG are often unresolved, and since NAA is the dominant metabolite, NAA is used interchangeably with NA for practical purposes.) Spectra showing only *mild* elevation of Cho or Cho/Cr are less helpful, because nonneoplastic conditions, including post–radiation therapy necrosis, can produce similar changes. In such cases, physiologic MR imaging that displays diffusion or perfusion properties can aid in narrowing the differential diagnosis. In general, the larger the increase in Cho or Cho/Cr, the more likely that a solitary focal or infiltrating lesion is a glial neoplasm when there is no history or other evidence of metastatic disease or hematologic malignancy. The converse of this statement may also be true; a Cho level or Cho/Cr ratio that is not elevated is unlikely to be a neoplasm. Nonetheless, as illustrated by Londono et al and Saraf-Lavi et al, the pitfall of excluding neoplasm from the differential diagnosis can and should be avoided. The key is to pay attention to the metabolite concentrations (Cho, Cr, and NA) and the ratios (low Cho/Cr values may be due to elevated Cr), and to evaluate other resonances, especially those at 3.5–3.6 (or 3.5 – 3.7) ppm and 3.35 ppm.

The report by Londono and colleagues is a sequel to an earlier study by the same group (1) in which patients with previously treated astrocytic neoplasms (low-grade astrocytoma, anaplastic astrocytoma, and glioblastoma multiforme [GBM]) were examined by short TE (stimulated echo acquisition mode) single voxel spectroscopy. In that study, Cho/Cr was elevated in the tumor spectra and showed a trend toward higher values in GBMs compared with those in low-grade astrocytomas, as has been generally reported; however, the intriguing finding was that values for the ratio of the 3.5–3.6 ppm resonance (presumed m-Ins) relative to Cr showed an inverse trend, being higher in low-grade astrocytomas than in GBMs. In the current case report, the finding of an elevated 3.5–3.6 ppm resonance in the multi-voxel spectra (spectroscopic imaging, using point resolved spectroscopy technique with a TE of 30 ms) from a pathologically proven low-grade astrocytoma (grade II according to the World Health Organization classification system) corroborates the earlier results, yet is accompanied by a new finding—no significant increase in Cho or in

Cho/Cr. Interestingly, this unusual combination of MRS findings corresponds to the findings in the case of gliomatosis cerebri (GC), reported by Saraf-Lavi et al in this issue. Thus, the reader may conclude that low-grade astrocytoma can have spectral findings similar to those of GC (a diffusely infiltrative lesion involving at least two lobes of the brain by definition), just as the former may resemble GC on routine MR imaging and stereotactic biopsy results. This conclusion, however, seems to contradict the recent hypothesis by Galanaud et al (2) that MRS can differentiate GC from low-grade glioma. Is Galanaud's hypothesis oversimplified? Perhaps. An alternative interpretation would be that the grade II astrocytoma reported by Londono et al is actually a case of GC! The spectral evidence necessary to confirm this latter interpretation is inconclusive, however, since Londono et al used different spectral analysis techniques from those used by Galanaud and colleagues to establish their criteria.

The relationships between spectral results and the pathophysiology of the tumors are discussed in both case reports. Regarding Cho and Cho/Cr, both groups of investigators attribute the relatively normal values to diminished (or disrupted) membrane lipid turnover and a low proportion of rapidly dividing cells ("lack of cellular proliferation") occurring in low-grade as compared with high-grade tumors. Tumor grade alone, however, seems unlikely to account for the Cho/Cr results because 1) the GC lesion was not strictly low grade and 2) most low-grade astrocytomas evaluated previously by Castillo and colleagues were found to have elevated Cho/Cr values (1). Similar issues have been raised by Galanaud et al. Londono and colleagues have suggested an alternative possibility: the presence of mixed oligodendroglial components in the tumor could have contributed to the unusual spectral results.

Regarding the 3.5–3.6 ppm resonance, its identity and the implications for pathophysiology are approached somewhat differently in the two reports. Two metabolites, m-Ins and Gly, have resonances detectable on short TE spectra in the 3.5–3.6 ppm region, and in healthy volunteers a single unresolved peak at approximately 3.56 ppm is usually assigned to m-Ins with a small contribution (~10%) from Gly. Londono and colleagues speculate that increased glycine, due to the presence of oligodendroglial cells, may account for the increase in the 3.5–3.6 ppm resonance. While there is some evidence in the literature to support this, there is also compelling in vitro MRS data suggesting that increased Gly occurs in high-grade (eg, GBM) rather than low-grade tumors (3). As discussed by Saraf-Lavi and colleagues, and by others (2, 4), a rough estimate of the Gly contribution to the 3.5–3.6 ppm resonance at short TE can be made by examining the same chemical shift region on long TE spectra. A strong resonance in this region would favor abundant Gly, since its -CH_2 group has a relatively long T2 and no J-coupling effects. m-Ins, on the other hand, produces a weak or absent resonance at TE 135 ms due to complex coupling of its –CH groups. This latter condition was observed by Saraf-Lavi and colleagues, who concluded that m-Ins primarily accounts for the increased 3.5–3.6 ppm resonance on short TE spectra.

Additional support for this conclusion may come from the prominent singlet resonance that was observed at ~3.35 ppm in both the short TE and longer TE GC spectra, consistent with elevated levels of s-Ins. m-Ins and s-Ins are the two most abundant isomers of the five naturally occurring stereoisomers of inositol, an amino alcohol. The concentration of s-Ins seems to be tightly coupled to the m-Ins concentration at a ratio of {12 m-Ins:1 s-Ins} (5), so that the elevated s-Ins might provide indirect evidence of elevated m-Ins. Unfortunately, this reasoning is oversimplified and possibly misleading because 1) the previously reported tight coupling was based on spectra from normal brain and not low-grade astrocytoma or GC and 2) even for normal brain the tight coupling has been questioned (6). Clearly the observations by Saraf-Lavi and colleagues on the prominent s-Ins peak deserve further study before their clinical significance can be assessed.

How does one explain the elevated m-Ins in low-grade astrocytomas? Two explanations, based on evidence that m-Ins is a "glial marker," are discussed by the authors of the case reports in this issue. In the first explanation, the elevation is attributed to changes in the phospholipid composition or abundance of glial cell membranes or both. Castillo and colleagues (1) earlier proposed a more specific mechanism in which mitogen-influenced metabolism of phophatidyl inositol (PI) results in 1) increased PI synthesis and corresponding depletion of the MR-visible m-Ins pool in high-grade astrocytomas, and conversely, 2) decreased PI synthesis and corresponding elevation of MR-visible m-Ins in low-grade astrocytomas. Galanaud et al (2) have proposed that the elevated m-Ins in GC is related to "proliferation of glial elements or, more probably, activation of normal glia." In the second explanation, elevation of m-Ins is attributed to its action as an organic osmolyte, playing a major role in the volume and osmoregulation of astrocytes, although the nature of this role in low-grade astrocytomas (or GC) versus that in high-grade astrocytomas remains to be elucidated.

In summary, what have we learned from these two case reports? First, for hyperintense lesions without specific morphologic findings on fluid-attenuated inversion recovery images and minimal or no enhancement on postcontrast T1-weighted images, the lack of significant Cho or Cho/Cr elevation does not exclude the diagnosis of a primary glial neoplasm. Second, a prominent resonance at 3.5–3.6 ppm likely signifies increased m-Ins, and both short TE and long TE spectra should be acquired in order to better characterize this and other resonances. Third, if m-Ins, m-Ins/Cr, and/or m-Ins/NA are shown to be elevated, include low-grade astrocytoma and gliomatosis cerebri in the differential diagnosis. Differentiation be-

tween these two lesions may become possible in the future with a better understanding of their spectral properties and underlying pathophysiology.

BRIAN C. BOWEN
Editorial Board

References

1. Castillo M, Smith JK, Kwock L. **Correlation of myo-inositol levels and grading of cerebral astrocytomas.** *AJNR Am J Neuroradiol* 2000;21:1645–1649
2. Galanaud D, Chinot O, Nicoli F, et al. **Use of proton magnetic resonance spectroscopy of the brain to differentiate gliomatosis cerebri from low-grade glioma.** *J Neurosurg* 2003;98:269–276
3. Kinoshita Y, Kajiwara H, Yokota A, Koga Y. **Proton magnetic resonance spectroscopy of brain tumors: an in vitro study.** *Neurosurgery* 1994;35:606–614
4. Mader I, Roser W, Hagberg G, et al. **Proton chemical shift imaging, metabolic maps, and single voxel spectroscopy of glial brain tumors.** *MAGMA* 1996;4:139–150
5. Michaelis T, Helms G, Merboldt KD, et al. **Identification of Scyllo-inositol in proton NMR spectra of human brain in vivo.** *NMR Biomed* 1993;6:105–109
6. Seaquist ER, Gruetter R. **Identification of a high concentration of scyllo-inositol in the brain of a healthy human subject using 1H- and 13C-NMR.** *Magn Reson Med* 1998;39:313–316

Measuring the Effect of Novel Therapies for Back Pain

A variety of ingenious new therapies have been introduced for the purpose of relieving back pain. Often, the early reports describing these therapies suggest substantial clinical benefit that is not confirmed by later studies. This occurrence poses questions: What are the hallmarks of a study that will have lasting scientific validity? What are the design features of a good clinical trial? Evaluating the benefit of therapies for back pain has complexities not found in all clinical trials, and specialized strategies are therefore needed for such studies. How can studies be structured to objectively assess a novel treatment of back pain?

Scientific validity is obtained in a clinical trial only when all sources of significant bias have been eliminated or minimized. Bias can be defined as any systematic error due to the design or conduct of a study. In any study of efficacy, biases may occur because of the way patients are selected, treatment is administered, or outcome data are acquired. Investigators in clinical trials must consider these potential sources of bias and must design their studies to minimize them. The Bellwether method is the prospective randomized trial in which outcomes are compared for patients randomized prospectively to treatment groups. Although this strategy minimizes enrollment bias by equilibrating potential confounding effects between groups, it is not the only method for obtaining objective results.

Many of the published studies of back pain therapies can be characterized as cohort studies, in which a series of patients subjected to the same therapy is evaluated. This is a study design without randomization and without a comparison group. A cohort study can provide useful information regarding the costs, complications, and selection criteria for therapy. The problem with a cohort study is that the effects of patient selection, treatment, and natural history of the disorder cannot be distinguished. Therefore, the cohort study is unreliable for measuring the therapeutic effect of a new treatment. Cohort studies of back pain therapies are numerous. Many of them show success rates greater than 80% for treatments such as spinal manipulation, epidural blocks, exercise, and even a placebo procedure (1). In most instances, more rigorous studies fail to confirm the high success rates of the early cohort.

How can high success rates be observed after intervention with ineffective procedures? Positive outcomes in a cohort may be explained by the tendency of patients with persistent symptoms to seek treatment at a time when their symptoms are most severe. For any condition that fluctuates in severity, when patients are monitored after therapy, they tend to have less severe symptoms than when they seek treatment. This tendency is an example of the statistical rule of regression toward the mean. In patients with back pain, fluctuation in severity is the rule. Because of regression toward the mean, patients with back pain treated with a placebo may seem to improve.

To determine the effect of treatment, treated patients must be compared with differently treated or untreated patients. Comparing two differently treated cohorts is one possible study design; however, a comparison of cohorts without the use of randomization may lead to biases that significantly undermine scientific validity. For example, cohorts from different clinics or different institutions may differ in sex, previous treatment, age, or stage at which treatment is initiated. The differences may be recognized by the investigators or may go unrecognized. In general, in retrospective cohort studies, equivalence between the cohorts is not feasible (2). Therefore, the cohort study generally has biases not found in prospective, randomized, controlled trials. The results of blinded randomized controlled trials are more reproducible than the results of retrospective cohort studies (2).

The appropriate comparison treatment to use in a controlled randomized trial of a procedure to treat back pain requires some judicious choices. Although a sham treatment may be scientifically correct, it may be ethically questionable. In a study of back pain, the control might be no treatment or treatment with another procedure. In a benchmark study to evaluate the benefits of lumbar diskectomy for the treatment of disk herniation, Weber (3) evaluated a series of patients, selected those who were considered candidates for diskectomy, and then assigned these pa-

tients on the basis of a randomization strategy to surgical treatment or to control treatment without surgical therapy. This study helped to measure the effectiveness of lumbar diskectomy and to establish the natural history of disk herniation. It is not, however, an easy study design to replicate.

Enrollment bias may be the most common flaw in cohort studies (4). If two groups to be compared differ in their initial state (ie, if one group has a factor that results in a better prognosis), the differences in outcomes may reflect these patient factors rather than differences in therapeutic efficacy. Therefore, enrollment criteria should be as concrete and objective as possible to minimize bias. Even the best enrollment criteria usually allow some flexibility. Even with rigid enrollment criteria, bias may still be present if the physicians enrolling patients for one treatment group apply the criteria less stringently for one group than for another treatment group. Ideally, investigators should not themselves have the responsibility to assign patients to the treatment or control group. Randomization of patients to the different treatment groups more effectively eliminates enrollment bias. Randomization may not suffice in every case. If variation in enrollment strategies across physicians is suspected in a randomized trial, physicians can stratify randomization. To minimize enrollment bias, patients should first be enrolled on the basis of explicit criteria and then be assigned, by a randomization strategy, to a treatment or control group.

Enrollment imbalances may be the cause of flawed or misleading results. An example is a randomized trial in which two groups were identified within the placebo-treated group who had significantly different mortality rates. The one group that took the placebo regularly had a different mortality rate than did the group that took the placebo irregularly. Because the mortality rate differences for these two groups of placebo takers cannot be attributed to differences in the placebo dose, the different outcomes must be explained by other factors, which are associated with whether a patient was compliant (1). One may attempt to adjust for the potential biases by controlling for the baseline status of the study participants. In back pain studies, previous surgery, compensation, duration of complaints, and psychologic makeup may be important predictors of outcome (5).

Even in randomized trials, enrollment biases may be present if the study samples are insufficiently large to assure an even distribution of the confounding factors. Therefore, even in a randomized trial, the clinical population must be analyzed in sufficient detail that the possibility of such errors is minimized. Differences in baseline conditions of the patients in the different treatment groups are dealt with by stratification. If the treatment groups are subdivided (stratified) by previous treatment, age, sex, and other factors that are known to be relevant, possible confounds may be identified. In cohort studies, possible enrollment biases may be dealt with by matching individual patients in one cohort to individual patients in another cohort who have the same age, sex, and other relevant features. Multicenter trials raise additional problems for the elimination of differences between treatment and control groups that require additional strategies (6). A detailed description of the enrollment criteria and randomization methods in a scientific report suggests that the investigators have seriously considered the confounding effects of possible enrollment bias.

Multivariate analysis of patient outcomes, although not a substitute for randomization, is a helpful tool to distinguish possible biases and adjust results accordingly. Adjustment for potential biasing factors is absolutely essential in cohort trials, in which there is no randomization. The possible differences between groups may be adjusted for by multivariate regression modeling. To adjust for the potential biases from differences in initial pain severity, previous treatment, overall health status, sex, and age, one could fit a regression model (linear for continuous data and logistic for binary), including these factors as well as treatment as covariates. The effect of treatment from this model would be known as an "adjusted effect." The effect of treatment that ignores the other factors would be known as "unadjusted." A comparison of a simple unadjusted analysis with the multivariate analysis will provide some indication regarding the confounding factors in the cohorts.

Enrollment criteria for patients undergoing experimental back pain treatments require special consideration. Individual variation may be especially marked among patients with back pain and disk degeneration. Because of this diversity among patients with back pain, the enrollment criteria must be thoughtfully defined to assure the appropriateness of the group for treatment. Selecting patients because of particular disk morphology is generally less acceptable than selecting patients with specific clinical signs and symptoms, because the relationship between back pain and disk degeneration is complex (7). The use of signs and symptoms, preferably in some standardized manner such as the Quebec Task Force classification, is preferable to the use of morphologic features of the spine (8).

Another major source of bias is the observer or detection bias. Different people assessing clinical outcomes at different times or employing different methods may observe different outcomes. Assuring objectivity and reproducibility in the measurement of outcomes is a major consideration in study design. Assessment of outcomes by investigators who have a stake in the research and who use subjective ratings is the least valid method. Subjective rating with the usual four-step scale (excellent, good, fair, poor) may produce very divergent results in the hands of different investigators (9). The most reliable outcome measures are those that are objective and based on physiology or activity. For example, in studies of back pain, a good question to ask in an outcome questionnaire would be, "Do you use pain medication more or less frequently since your procedure?" and a bad question would be, "Do you feel better or worse since you began treatment?"

To minimize observer bias, various strategies can be used. Double blinding, in which both the physician

and the patient are unaware which treatment has been applied, is the best method to avoid observer bias but the most difficult to implement in back pain therapy studies. Double blinding eliminates the possibility that the patient or the doctor manipulates the treatment or biases the reporting. In cohort and retrospective studies, double blinding is virtually impossible. For studies of back pain treatment, the choice of outcomes to measure is challenging because of the variety of signs and symptoms and temporal fluctuations. In cases of back pain, outcomes are typically dissociated; one measure improves while another deteriorates. Outcomes for back pain studies cannot be measured in terms of laboratory or imaging measurements but must rely on patient experience, considering the nature of back disease. Methods for measuring outcomes of this sort are valid and reliable (10). In a clinical trial, a detailed description of the methods of assessing patient outcomes and selection of observers who are independent of the therapeutic team reassures readers that the investigators have attempted to achieve objectivity in their study.

Another potential systematic error is the transfer bias, which is a differential rate of attrition in the treated compared with the control group (11). Complete follow-up is not achieved in all cases. Transfer bias may confound results in that hostile patients and noncompliant or poorly responding patients may be lost to follow-up to a greater degree than other patients. In clinical trials, it cannot be assumed that the success rate is the same in the cases lost to follow-up as in the cases followed. A high attrition rate may suggest a potential bias. The number of study participants lost to follow-up should be reported in any clinical trial. Analyses to evaluate a possible transfer bias should be included in clinical trials. Adjustments may be made or post hoc matching may be used to minimize this source of bias.

Study designs without controls and randomization may answer questions of clinical importance regarding novel treatments for back pain (11), but the randomized clinical trial remains the criterion standard for assessing clinical efficacy. Good examples of randomized, controlled, prospective trials of surgical and nonsurgical treatments for back pain are found in the literature (12, 13). The technical and ethical concerns regarding randomization can usually be overcome by one or another randomization scheme (11). If randomized clinical trials are not feasible for ethical, economic, or logistical reasons and if a cohort design is chosen, the investigators should attempt to achieve the highest degree of scientific validity possible. Matching or adjusted multivariate regression analyses, and other methods, may be used effectively. The performance of clinical trials is difficult and demands the utmost honesty and neutrality from the researcher (7). High standards of investigation, similar to those for conducting clinical trials for cancer or cardiovascular disease, applied to the evaluation of back pain therapies would benefit both patients and physicians who are interested in knowing the benefits of spinal therapies.

Can randomized controlled study designs be used to measure the therapeutic efficacy of intradiskal therapies, such as the one described in this issue of the *AJNR*? Patients seeking treatment for back pain in a medical facility may be asked to participate in a study and informed that their treatment will be randomized. Those who agree to participate will be screened further. Those meeting inclusion and exclusion criteria of duration, character, and severity of pain who are considered candidates for the experimental treatment are then randomly assigned to the experimental treatment or other treatment(s) (eg, treatment with intradiskal medical ozone, treatment with intradiskal steroids, or treatment with a combination of the two). As a minimum, the treating physicians must not participate in the assignment of patients to treatment groups. Blinding the treating physician to the treatment administered might be an additional option to minimize bias. For example, a technician who alone knows the randomization plan and keeps a log hidden to the other investigators might provide the therapist with the material to inject into the disk. The patient may be blinded to the treatment by not informing him or her which therapy was used. An independent observer, blinded to the treatment group of the patient, evaluates each patient at a specified time point or time points after treatment according to a predetermined and validated instrument such as a questionnaire. Objective measures of a patient's condition, such as number of analgesics used per day, are included. Ideally, the same observer evaluates each patient or, if this is not feasible, the selected observers are randomly assigned to evaluate patients. The outcomes of the patients assigned to each group are compared with statistical tests. By convention, outcomes are reported in terms of the patients assigned to each group (intention to treat) rather than the number receiving treatment or the number receiving technically sufficient treatment. The effect of technical failures may be analyzed and discussed. Other possible confounding factors are detected, analyzed, and discussed. Patient features such as age, duration of symptoms, and sex in the three groups are tested for differences and their effects on results considered. Drop out rates are recorded and their possible effect analyzed. If statistically significant differences are found between groups, without evidence of important biases, conclusions may be tentatively drawn concerning the relative efficacy of medical ozone therapy versus steroid therapy alone or a combination of the two.

VICTOR M. HAUGHTON
Editorial Board
JASON FINE
University of Wisconsin
Madison, WI

References

1. Deyo R. **Practice variations, treatment fads, rising disability: do we need a new clinical research paradigm?** *Spine* 1993;18:2153–2162
2. Hoffman RM, Turner JA, Cherkin DC, Deyo RA, Herron LD. **Therapeutic trials for low back pain.** *Spine* 1994;19:2068S–2075S

3. Weber H. **Lumbar disk herniation: a controlled, prospective study with ten years of observation.** *Spine* 1983;8:131–140
4. Keller RB, Rudicel SA, Liang MH. **Outcomes research in orthopaedics.** *Instr Course Lect* 1994;43:599–611
5. Deyo RA. **Measuring functional status of patients with low back pain.** *Arch Phys Med Rehabil* 1988;69:1044–1053
6. Revel M, Payan C, Vallee C, et al. **Automated percutaneous lumbar discectomy versus chemonucleolysis in the treatment of sciatica: a randomized multicenter trial.** *Spine* 1993;18:1–7
7. Weber H. **The natural history of disc herniation and the influence of intervention.** *Spine* 1994;19:2234–2238
8. Loisel P, Vachon B, Lemaire J, et al. **Discriminative and predictive validity assessment of the Quebec Task Force classification.** *Spine* 2002;27:851–857
9. [No authors listed]. **Scientific approach to the assessment and management of activity-related spinal disorders: a monograph for clinicians: report of the Quebec Task Force on Spinal Disorders.** *Spine* 1987;12[suppl]:S1–S59
10. Deyo RA, Diehl AK. **Psychosocial predictors of disability in patients with low back pain.** *J Rheumatol* 1988;15:1557–1564
11. Keller RB. **Pro: outcomes research is cost effective and critical to the specialty.** *Spine* 1995;20:384–386
12. Rasmussen FO, Amundsen T, Vandvik B. **Lumbar disk prolapses and radiologic spinal intervention: what do the randomized controlled trials say? [in Norwegian]** *Tidsskr Nor Laegeforen* 1998;118:2470–2480
13. Burton AK, Tillotson KM, Cleary J. **Single-blind randomised controlled trial of chemonucleolysis and manipulation in the treatment of symptomatic lumbar disc herniation.** *Eur Spine J* 2000;9:202–207