MR Contrast Media in Neuroimaging: A Critical Review of the Literature

Jonathan Breslau, Jeffrey G. Jarvik, David R. Haynor, W. T. Longstreth, Jr, Daniel L. Kent, and Kenneth R. Maravilla

BACKGROUND AND PURPOSE: MR contrast media are commonly used but do not have evidence-based guidelines for their application. This investigation seeks to define specific methodological problems in the MR contrast media literature and to suggest guidelines for an improved study design.

METHODS: To evaluate the reported clinical efficacy of MR contrast media in neuroimaging, we performed a critical review of the literature. From 728 clinical studies retrieved via MEDLINE, we identified 108 articles that evaluated contrast media efficacy for a minimum of 20 patients per study. The articles were randomly assigned to four readers (a fifth reader reviewed all of the articles) who were blinded to article titles, authors, institutions, and journals of publication. The readers applied objective, well-established methodological criteria to assign each article a rating of A, B, C, or D.

RESULTS: One hundred one of 108 articles received a D rating, six received a C rating, and one received a B rating. In general, the Methods sections of the evaluated articles did not contain details that would allow the reader to calculate reliable measures of diagnostic accuracy, such as sensitivity and specificity. Specifically, a common problem was failure to establish and uniformly apply an acceptable standard of reference. In addition, images were not always interpreted independently from the reference standard. Radiologists and clinicians need to determine the applicability of any published study to their own practices. Unfortunately, the studies we reviewed commonly lacked clear descriptions of patient demographics, the spectrum of symptomatology, and the procedure for assembling the study cohort. Finally, small sample sizes with inadequate controls were presented in almost all of the articles.

CONCLUSION: Although MR contrast media are widely used and play an essential role in lesion detection and confidence of interpretation, no rigorous studies exist to establish valid sensitivity and specificity estimates for their application. On the basis of this review, we herein describe basic methods to document improvements in technology. Such studies are essential to devise measures of diagnostic accuracy, which can form the basis for further studies that will assess diagnostic and therapeutic impact and, ultimately, patient outcomes.

Gadolinium-based MR contrast agents have been widely applied since they were first available for clinical use in 1988. By March 1993, more than 5.4 million doses had been administered (1). At an approximate hospital charge of $150 per dose, the use of contrast material accounted for almost $1 billion during the first 6 years of its clinical application. Most contrast material was used in neuroimaging, with an established role in lesion detection, in characterization, and in improving radiologists’ confidence regarding interpretation.

In the current environment of cost containment, an established role may not warrant continued government and corporate support for a specific technology. Increasingly, these financially involved entities require evidence-based practice guidelines and patient outcomes data based on rigorous technology assessment methodology. We sought to apply such criteria to the evidence for clinical efficacy; that is, to the probability that a patient will derive benefit, under optimum conditions, from
contrast-enhanced MR neuroimaging. This analysis could yield evidence-based guidelines for the use of contrast material or serve as a demonstration of what is lacking in the evidence at this time.

**Methods**

**Article Selection**

We undertook a comprehensive literature search using the MEDLINE database, merging the subject headings “magnetic resonance imaging” and “contrast media” and the key words “nervous” or “brain” or “spine.” Limiting our search to the English literature until mid-1997, we found 728 clinical studies reporting on the use of contrast-based MR contrast media in neuroimaging. After excluding case reports, reviews, and articles reporting on fewer than 20 patients, we found a total of 108 articles evaluating the efficacy of MR contrast media. These 108 articles were randomly distributed among four readers, who were blinded to the titles, authors, institutions of origin, and journals of publication. A fifth blinded reader evaluated all the articles. Each qualifying article was thus rated twice, with disagreements resolved by discussion and consensus.

**Rating Criteria**

The articles were assigned a rating of high, intermediate, or low for each of seven well-established criteria. The criteria were previously published in technology assessment articles (2).

The **Technical Quality of the Index Test** rates the technical quality of the MR equipment and contrast dosage used in a study. High-quality articles reported the administration of at least a 0.1-mmol/kg dose of contrast agent with a magnet field strength of at least 1.0 T and a section thickness of not greater than 5 mm. Intermediate-quality articles had medium field (>0.3 T) magnets or large or unspecified section thickness. Low-quality articles had low-field magnets or did not specify the index test quality.

The **Technical Quality of the Reference Test** addresses the quality of the standard of reference applied in a study. High-quality reference tests included pathologic proof, surgical findings, or comprehensive clinical follow-up, with specific criteria for establishing the particular diagnosis. A rating of intermediate was assigned if the diagnostic criteria were incompletely defined. If the standard of reference was undefined, studies were rated low. A low rating was also assigned if no tests beyond the index test were applied.

The **Application of the Reference Test** evaluates the thoroughness with which the standard of reference was applied. A study was considered to be of high quality if it used the same reference test for all cases and intermediate studies used different but all acceptable, reference tests. Low-quality studies did not use an acceptable reference standard for all cases. If no standard of reference existed, articles were rated low for both reference test quality and application.

The **Independence of Interpretation** assesses the separation maintained between interpretation of the index and reference tests. If these two tests or standards were explicitly interpreted independently of each other, a rating of high was assigned. If either test review bias (lack of blinding to the final diagnosis when interpreting the index test) or diagnosis review bias (index test result influencing final diagnosis) was present, the article was rated intermediate. If both biases were present or if information was not available regarding independence of interpretation, an article was rated low for this criterion.

The **Clinical Description** refers to the detail with which the patients’ clinical presentations were described. High-quality studies contained thorough clinical and demographic information that included at least age, sex, and percentage of patients displaying major relevant signs and symptoms. Studies with incomplete descriptions were considered to be of intermediate quality. When the description was limited (e.g., “suspected intracranial pathology”) or nonexistent, an article was rated low.

The **Cohort Assembly** refers to the methods used to select cases for a study. A quality rating of high for cohort assembly required prospective enrollment from a primary care setting, with a range of clinical presentations. A cohort assembled in this fashion would have relatively little referral filtering. Alternatively, investigations of diseases generally not encountered in the primary care setting, such as complex partial seizures, could be rated high for cohort assembly as long as cases were accrued prospectively without workup bias. If cases were selected retrospectively from referral centers or if the fact of testing was the criterion for enrollment, the study was considered to be of intermediate quality. Finally, cases with workup bias (selection because of a positive index test result) or no description of cohort assembly were rated low.

The **Sample Size** refers to the number of cases and control subjects included. High-quality studies had at least 35 cases and 35 control subjects. A sample size of 35 is the minimum for which the lower bound of the 95% confidence interval for a true sensitivity or specificity of 1.0 would exceed 0.9. If a study contained either fewer than 35 patients or fewer than 35 control participants, it was rated intermediate. If a study had both fewer than 35 patients and fewer than 35 control subjects, it was rated low.

A summary rating was assigned to each article based on the seven quality criteria (2, 3). An article was rated A if it was at least intermediate in clinical description but high in all other criteria. B articles were high in reference tests and independence but could be intermediate in all other categories. C articles could be low in clinical description but had to have at least intermediate ratings in all other categories. Articles not meeting C criteria received a D rating.

**Results**

Of a total of 756 ratings (seven criteria for 108 articles), there were 50 disagreements (7%). The disagreements were most commonly regarding the ratings of the reference test, cohort assembly, and clinical description.

Of the 108 articles rated, one was rated B (4), six were rated C, and 101 were rated D. The distribution of ratings among the seven quality criteria is listed separately (see the table). Index test quality received the highest ratings (61 of 108) because the reporting institutions generally used state-of-the-art equipment. The authors focused most of the attention on the quality of imaging. Reference test quality also frequently earned high or intermediate ratings (33 of 108 high) because of the availability of pathologic specimens or close surgical collaboration. The availability of pathologic proof often was the means of selection into a study. The difficulty in uniformly applying an acceptable standard of reference led to lower quality ratings in the application of the reference test. Authors failed to apply stringently the requirement that all analyzed cases have the same reference standard. In some articles, a few cases with no standard of reference could have been eliminated while maintaining an adequate sample size. Approximately one third of the articles did not use a reference standard. These shortcomings precluded the calculation of accuracy
statistics because positivity and negativity of the contrast-enhanced MR imaging could not be assigned.

Regarding the next four criteria, very few high ratings were assigned. Independence of interpretation, which rates the avoidance of review bias, is a central concern in study design. In general, the Methods section of articles did not document explicit separation between the interpretation of the contrast-enhanced MR imaging findings and the standard of reference. Intermediate articles usually described interpretation of the contrast-enhanced MR images without knowledge of the final diagnosis but did not describe the procedures used to prevent the MR findings from affecting the final diagnosis. The assessment of clinical description yielded only 11 high ratings. In many articles, the clinical description consisted only of summary statements, such as “suspected intracranial pathology.” It is likely that many studies could have been improved by including reviews of medical record. On the other hand, cohort assembly represents the most difficult aspect of study design for radiologists, who may have little influence over the spectrum of disease they see. Correspondingly, only six articles (5–10) were rated high for cohort assembly. Unless they have close clinician collaboration from the beginning of a study design, radiologists must focus on retrospective case selection. In addition, case accrual at a tertiary care center usually includes substantial referral filtering. Finally, almost all articles presented sample sizes that were inadequate to yield robust statistics and did not present control cases. Not one of the five articles with high ratings for sample size (5, 11–14) applied reference tests.

### Discussion

This critical literature review, consisting of structured blinded ratings, was originally designed as a metaanalysis, to derive pooled estimates of sensitivity and specificity for the use of contrast material in neuroimaging. The available studies, unfortunately, did not themselves yield valid accuracy measures, which precluded metaanalysis. One finding of our study, therefore, is that no valid measures of sensitivity or specificity exist for the application of MR contrast material in neuroimaging.

We did apply our study design to uncover surprisingly prevalent methodological flaws. In a critical literature review evaluating diagnostic tests in general during a 16-year period. Reid et al (15) similarly found inadequate assessment of diagnostic tests, although use of methodological standards did increase during the study interval from 1978 to 1993. In their study, radiologic tests were the largest single category of diagnostic test evaluated. Although their results were described as particularly disturbing, our results document a much lower prevalence of acceptable methodological standards. For example, Reid et al reported that 47% of analyzed articles avoided review bias during the interval between 1990 and 1993, as compared with only 4% (four of 108) in our study.

The one article that was assigned a B rating contains basic methodological elements that any study of diagnostic accuracy should include. In 1992, Wiebe et al (4) reported their study using craniospinal MR imaging with contrast enhancement to serially examine patients with multiple sclerosis (MS). All patients underwent cranial and spinal MR imaging on at least three occasions at 13-week intervals, with additional imaging performed if clinical relapses occurred during the study period. The clinical judgment of a neurologist at a university MS clinic with respect to presence or absence of disease activity was used as the standard of reference. Patients with quiescent disease were included in the study. The examining neurologists were blinded to the MR findings, and the radiologists were blinded to the patients’ clinical status. The article contains a table that clearly describes the spectrum of disease evaluated. In this manner, the authors were able to construct a standard two-by-two table showing test positivity and negativity, as well as the presence or absence of disease. The Methods section presents details that allow the reader to recognize the limitations in generalizing, which include a somewhat narrow spectrum of disease and a small sample size.

The reported accuracy statistics in the article by Wiebe et al (4), although favorable, were for the use of MR imaging overall. Contrast enhancement
for cranial imaging was used only in those cases in which evidence of activity had already been noted on unenhanced sequences. However, contrast enhancement was the sole evidence of disease activity in 5% of all spinal cord images. These results parallel the clinical use of contrast material in the evaluation of patients with MS.

Because low-field MR systems are gaining acceptance, we undertook an informal analysis of our data and assigned high ratings to all articles for index test quality. We found that no article gained a higher overall rating, because the rating system values the presence and quality of a standard of reference over other criteria.

Conclusion

We have shown that strong evidence-based guidelines for the use of contrast material in neuroimaging cannot be derived from the current literature. In short, the clinical efficacy of contrast material remains unproved. Future investigations need to focus on constructing robust measures of diagnostic accuracy. These studies will require larger sources of funding to construct appropriate randomized controlled trials, as demonstrated by the highest rated article in our study. Researchers in neuroradiology need to apply such methodologies, which have been used successfully in advancing other areas of medicine. Only when we rigorously evaluate the ability of contrast material to aid in the diagnosis and exclusion of disease can we proceed to the evaluation of its diagnostic and therapeutically relevant impact.

Appendix

Articles reviewed:
Carvlin MJ, DeSimone DN, Meeks MJ. Phase II clinical trial of gadoteridol injection, a low-osmolar magnetic resonance imaging contrast agent. Invest Radiol 1992;27:S16–S21
Deutsch AL, Howard M, Dawson EG. Lumbar spine following successful surgical discectomy: magnetic resonance imaging features and implications. Spine 1993;18:1054–1060


Runge VM, Bronen RA, Davis KR. Efficacy of gadoteridol for magnetic resonance imaging of the brain and spine. *Invest Radiol* 1992;27:525–532


References


