Reproducibility of Detecting Silent Cerebral Infarcts in Pediatric Sickle Cell Anemia

Robert I. Liem, MD, MS[1], **Jingxia Liu, PhD**[2], **Mae O. Gordon, PhD**[2], **Bruce A. Vendt, BS, MBA**[3], **Robert C. McKinstry III, MD, PhD**[4], **Michael A. Kraut, MD, PhD**[5], **John J. Strouse, MD, PhD**[6], **William S. Ball, MD**[7], and **Michael R. DeBaun, MD, MPH**[8]

[1]Division of Hematology, Oncology & Stem Cell Transplant, Ann & Robert H. Lurie Children's Hospital of Chicago, Chicago, IL

[2]Division of Biostatistics, Washington University School of Medicine, Saint Louis, MO

[3]Electronic Radiology Lab, Mallinckrodt Institute of Radiology, Washington University School of Medicine, Saint Louis, MO

[4]Department of Pediatrics, Washington University School of Medicine, Saint Louis, MO

[5]Divisions of Neuroradiology, Johns Hopkins Hospital, Baltimore, MD

[6]Pediatric Hematology, Johns Hopkins Hospital, Baltimore, MD

[7]Division of Pediatric Neuroradiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH

[8]Division of Pediatric Hematology/Oncology, Vanderbilt University Medical Center, Nashville, TN

Abstract

Detecting silent cerebral infarcts (SCI) on MRI in children with sickle cell anemia (SCA) is challenging, yet reproducibility of readings has not been examined in this population. We evaluated consensus rating, inter- and intra-grader agreement associated with detecting SCI on screening MRI in the Silent Infarct Transfusion (SIT) Trial. Three neuroradiologists provided consensus decisions for 1,073 MRIs. A random sample of 53 scans was re-analyzed in blinded fashion. Agreement between first and second consensus ratings was substantial ($\kappa = 0.70$, $p < 0.0001$), as was overall inter-grader agreement ($\kappa = 0.76$, $p < 0.0001$). In the test-retest sample,

intra-grader agreement ranged from κ of 0.57 to 0.76. Consensus decisions were more concordant when MRIs contained more than one lesion and lesions were larger. We conclude that the routine use of MRI to screen for SCI in the research setting is reproducible in SCA and agreement among neuroradiologists is sufficient.

Keywords

Sickle cell disease; stroke; reliability testing

INTRODUCTION

Cerebrovascular disease among individuals with sickle cell anemia (SCA) has a spectrum of presentation ranging from acute, neurologically devastating infarction to clinically silent cerebral infarction. First time clinically overt strokes, which are associated with focal neurological dysfunction, occur in up to 11% of unscreened individuals with SCA by age 20 years old.[1] Overt strokes in children with SCA typically occur in the distribution of large vessels, such as the distal internal carotid or middle cerebral arteries. In contrast, silent cerebral infarcts (SCI) are not associated with acute neurological symptoms and are distinguished from overt strokes by smaller lesions frequently in the deep cortical white matter.[2, 3] Although SCI have been noted in 27% of children with SCA less than 6 years old,[4] no consensus exists regarding their treatment.

Despite the recognition that SCI may be progressive[3] and associated with an increased incidence rate of overt strokes,[5] neurocognitive deficits,[6, 7] and poor academic attainment,[8] routine screening to detect SCI in children with SCA has generally been limited to pediatric hematology centers with a clinical and research focus on acquired brain injury in this population. The diagnosis of SCI may also be difficult given their smaller size when compared to overt strokes, coupled with the lack of clinical symptoms and the requirement for specific neuroradiology expertise available at limited larger pediatric centers.

SCI have previously been defined as foci of abnormally elevated T2-weighted signal intensity on MRI that measure at least 3 mm in one or more axes and are seen in two imaging planes.[9] The determination of SCI requires that focal neurological symptoms or signs that might correlate with lesion location be absent on history and physical examination.[10, 11] Few studies have examined reader agreement in the measurement of SCI or white matter lesions.[12–14] Based on lesion size alone, inter- and intra-observer agreement may be lower in individuals with SCI when compared to overt strokes.[15, 16] To our knowledge, no study has evaluated intra- and inter-observer agreement related to the detection of SCI in SCA or in children even though the clinical implications of such findings are considerable.

The Silent Infarct Transfusion (SIT) Trial is a multi-center, randomized clinical trial to evaluate the safety and benefit of chronic transfusion therapy in children with SCA and SCI. Study eligibility depends on accurate determination of the presence of SCI on screening MRI. Based upon clinical experience, the trial leadership assumed *a priori* that the detection of SCI on MRI would be associated with some degree of variability. As such, the trial

provides an opportunity to examine agreement among three neuroradiologists, who comprise the neuroradiology adjudication panel for the trial, to independently assess the presence or absence of SCI on MRI over time.

We hypothesized that a high level of agreement would be achieved among neuroradiologists faced with detecting SCI on screening MRI in children with SCA enrolled in the SIT Trial. The primary objective of this study is to describe the consistency of the consensus decisions made by the neuroradiology panel of the SIT Trial when screening MRIs were reinserted into their work queue without their knowledge. A secondary objective is to determine inter- and intra-rater agreement related to the detection of SCI on screening MRI obtained at study entry and on re-inserted studies, respectively. The last objective is to examine whether number, size and location of SCI affected consensus decisions on MRI adjudication.

METHODS

The SIT Trial relied on a centralized, digital workflow in which all MRIs are electronically transmitted from participating institutions for central review.[9] Three neuroradiologists at separate institutions comprised the neuroradiology adjudication panel for this trial and performed blinded assessments of each MRI. Patient and study identifiers were stripped prior to insertion of each assessment MRI into a digital queue for reading. Before the trial began, each of the three neuroradiologists underwent web-based training using a standard set of MRIs developed to instruct them on the detection of SCI. Together, these neuroradiologists established a working differential diagnosis for increased T2-weighted signal intensity on MRI specifically for the trial.

SIT Trial Screening MRIs

Each neuroradiologist independently established whether or not the technical quality of each screening MRI was sufficient for determination of the presence of a SCI. The presence or absence of a SCI was determined by independent responses ("yes", "no" or "indeterminate") by each neuroradiologist to the question "Based on the results of the MRI, the patient has at least one infarct-like lesion". Consensus decisions were concordant if all three neuroradiologists reached the same response. Disagreements were resolved by conference call to render a consensus decision (Figures 1 and 2). For each screening MRI that contained a SCI, these lesion characteristics were recorded: total lesion number, location of each lesion and volumetric size of each lesion. Each lesion was classified as located in either the right or left parietal, occipital, temporal or frontal lobe of the brain. Lesion location and size were evaluated only on studies with single lesions (N=88) to minimize confusion regarding which lesion or lesions contributed to a neuroradiologist's primary response.

Test-Retest Agreement Sample

A total of 53 MRIs for evaluating test-retest agreement were selected from a proportionate random sample of screening MRIs determined to be either positive or negative for the presence of a SCI by consensus. These scans constituted the quality control test-retest sample used to assess the consistency of the neuroradiology consensus decisions. Indeterminate scans were excluded from the test-retest sample since they comprised fewer

than 5% of the consensus decisions. Proxy clinic and patient IDs were assigned to the test-retest sample of screening MRIs, which were re-analyzed independently by the three neuroradiologists. Neuroradiologists were unaware of the re-insertion of the quality assurance MRIs and were masked to the initial consensus rating of each screening MRI in the test-retest sample.

Statistical considerations

Simple kappa (κ) statistic was calculated to reflect agreement between the first and second consensus decision by the neuroradiology panel. To calculate inter-grader agreement, κ statistics for ordinal data were calculated using the pre-consensus decisions of the three neuroradiologists, also referred to as graders. This overall inter-grader κ statistic was computed using the SAS Macro %MAGREE, and intra-class correlations (ICC) were computed using the SAS Macro %INTRACC. To assess intra-grader agreement between the first and second reviews of the same MRI study in the test-retest sample, weighted κ statistics were calculated for each grader. Weighted κ statistics were also computed for the agreement between each grader's ratings and the consensus decisions for both the entire screening set of MRIs and the test-retest sample. Jonckheere-Terpstra and Cochran-Armitage tests were used to test trends in κ statistics and positive MRI studies, respectively, over time. In accordance with previously accepted criteria by Landis and Koch, κ values greater than 0.4 indicate "moderate agreement", values greater than 0.6 indicate "substantial agreement" and values greater than 0.8 indicate "excellent" or "near perfect agreement".[17] Finally, logistic regression modeling and Chi-square/Fisher's exact test were used to assess the relationship of lesion and subject characteristics, both continuous and categorical, to consensus decisions. Kruskal-Wallis test was performed to examine the association between lesion size and consensus decisions for single lesions. All statistical analyses were performed using SAS, version 9.3 (SAS Institute Inc., Cary, NC, USA).

RESULTS

Demographic Characteristics of Screening and Test-Retest Samples

Of the total 1,091 screening MRIs subjected to quality control review, 1,073 scans were judged to be gradable and included in this study. A subset of 53 scans, which represented the test-retest sample, was randomly selected from the 1,073 screening MRIs and reassessed by all three neuroradiologists. For this test-retest sample, average time between the initial scan and its reinsertion into the work queue was 3 years with a range of 5 months to 5 years. The age of the subjects whose MRIs were included was similar in the screening versus test-retest samples (9.1±2.5 vs. 9.0±2.4 years) (Table 1). The proportion of males included in the two samples was also similar (50.9 vs. 54.7%). In both groups, the majority of subjects were of African descent and had hemoglobin SS as the primary genotype. Almost all of the MRIs in the screening and test-retest samples were obtained at 1.5 Tesla.

Consensus Agreement for Test-Retest MRI Sample

The overall agreement between the first consensus and second consensus ratings in the test-retest subset of screening MRIs was substantial (κ = 0.70, 95% CI [0.51, 0.90], $p < 0.0001$). A total of eight MRIs in this sample had discordant consensus ratings between the first and

second reviews. Of these, six MRI consensus ratings went from "no" on the first review to "yes" on the second review, and two ratings went from "yes" to "no". The demographic characteristics of the subjects and scans in this discordant group were similar to that of the screening and test-retest samples.

Inter-grader Agreement for All Screening MRIs

The overall individual agreement among the three neuroradiologists in determining if a SCI was present or absent on the 1,073 screening MRIs was substantial ($\kappa = 0.75$, $p < 0.0001$). Discordant ratings were observed among the three neuroradiologists for only 24 (2.2%) out of all screening MRIs. Agreement was strongest for the "yes" rating category ($\kappa = 0.80$, $p < 0.0001$). Substantial inter-grader agreement among the three graders was also supported by a calculated ICC Shrout-Fleiss reliability single score of 0.80.

Intra-grader Agreement for Test-Retest MRI Sample

MRIs in the test-retest sample were re-graded by all three neuroradiologists. A weighted κ statistic for intra-grader agreement between the first review and second review was calculated for each neuroradiologist. Intra-grader agreement for re-inserted MRIs ranged from a weighted κ of 0.57, 95% CI [0.35, 0.78] for Grader 1 to 0.76, 95% CI [0.56, 0.92] for Grader 2 (Table 2). For Grader 1, the number of "no" ratings on the first review that changed to "yes" on the second review was 9 out of 32 (28.1%), and the number of "yes" ratings that changed to "no" was 2 out of 19 (10.5%). The high percentage of changes in ratings explained why intra-grader agreement for Grader 1 was lower than that observed for the other 2 graders. We found that 10/53 (18.9%) re-inserted MRIs contributed to discordant ratings for at least 2 graders. No MRIs resulted in discordant ratings for all 3 graders.

Grader-Consensus Agreement over Time

Overall, the agreement between individual ratings and consensus ratings for screening MRIs was excellent for all three neuroradiologists, with weighted κ statistics ranging from 0.87 for Grader 2 to 0.91 for Grader 3 (Table 3). Each grader's individual rating was also compared to consensus ratings by year of enrollment in the SIT Trial from 2004 to 2010. Jonckheere-Terpstra testing demonstrated that grader-consensus agreement significantly improved for Graders 1 (test statistic = 1.69, $p = 0.045$) and 3 (test statistic = 2.07, $p = 0.019$), indicating a greater likelihood for larger weighted κ values over time, but remained relatively constant for Grader 2. However, the improvement in agreement for Graders 1 and 3 over time did not result in any significant change in the proportion of positive MRIs.

Factors Affecting Consensus Decisions

For screening MRIs with evidence for SCI, higher lesion number was significantly associated with greater concordance on consensus decisions (OR 2.13, 95% CI [1.51, 3.00], $p < 0.0001$). On MRIs with single lesions, lesions on average were greater in size for studies in which consensus decisions were concordant versus discordant (153.6 vs. 93.6 mm^3, $p = 0.055$), although this finding did not reach statistical significance. Neither lesion location nor subject characteristics such as age or gender significantly influenced whether consensus decisions were concordant or discordant on MRIs with a single lesion (data not shown).

DISCUSSION

Increasingly, SCI are recognized as a prevalent and progressive cerebrovascular complication in children with SCA,[18] but they may be challenging to detect on routine screening by MRI. This report represents the first to demonstrate that consensus detection of SCI on MRIs such as those obtained in the SIT Trial is reproducible in children with SCA. Although detecting SCI is clinically important, the routine use of MRI as a reliable tool for screening children with SCA for SCI has not been established. The overall inter-grader as well as intra-grader agreement in our study confirms the feasibility of detecting SCI on MRIs. Further, we show that the evaluation of SCI also remains consistent over time. A learning effect was evident in two of the three neuroradiologists, for whom improvements in individual grader versus consensus agreement were observed over time. Finally, we demonstrate that lesion number and size affect ability to detect SCI on screening MRI in this population.

Our assessment of the reproducibility associated with detecting SCI on screening MRI in the SIT Trial was rigorous for several reasons. A centralized, digital workflow allowed for the electronic transmission and insertion of MRIs stripped of patient identifiers to ensure blinded, independent assessments by each grader.[9] All three neuroradiologists underwent standardized training in the recognition of SCI, and the procedures for evaluating the presence or absence of lesions, including a working definition of SCI and their differential diagnosis, were outlined prior to the start of the study. Discrepancies in grader ratings for any MRI were subjected to adjudication and a consensus decision reached. MRIs that comprised the test-retest sample were re-inserted into the digital workflow in a masked fashion so that the neuroradiologists were unaware of which MRIs were "live" and which were quality control scans, also ensuring that intra-grader agreement would be evaluated without bias. There was low agreement between Grader 1's first and second reviews of the same MRIs because Grader 1 was more likely to find new lesions upon repeat review of the re-inserted scans. This may reflect learned behavior or an improvement in detection due to the adjudication process. Overall, a learned behavior associated with improved detection of SCI was also reflected in the consensus agreement for the test-retest sample as well as in grader-consensus agreement over time. Although there may be concern this learned behavior might result in bias toward more positive findings at the end of the SIT trial, there in fact was a stable trend, rather than an increase, in the proportion of positive MRIs over time.

Differences in patient characteristics, measurement techniques, rating scales, agreement testing approach and statistical methodology make it difficult to compare our results with that of other studies that have examined observer agreement in detecting white matter lesions.[12–14, 19] Subjects in previously published studies comprise mostly older individuals from large cohort studies at high risk for cerebrovascular disease. The pathophysiologic processes that underlie development of SCI in SCA versus deep white matter changes or lacunar infarcts in the general adult population are also potentially different. Nonetheless, inter-grader agreement in our report was comparable to that described in the Kapeller et al. study,[13] in which inter-grader agreement ranged from 0.599 to 0.781 for baseline MRIs, and in the Wardlaw et al. study,[14] in which intergrader agreement ranged from 0.720 to 0.890. Intra-grader agreement in our study was superior to that observed in the De Schryver et al.

study,[12] in which investigators evaluated 126 scans twice one month apart for the presence of white matter lesions ($\kappa = 0.580$). With the exception of lesion number, the influence of other lesion characteristics on grader consensus was not significant. Though no prior studies have examined the relationship between lesion number and detection on MRI, we anticipated there would greater agreement when more than one lesion was present.

Several limitations exist in this study. One is the relatively small size of our test-retest sample; however, our sample size was comparable to that of other similar studies published. The even smaller number of discordant ratings in the test-retest sample made it challenging to explore scan or subject characteristics that might have contributed to changes in ratings on a consensus or individual level. Possibly, our three neuroradiologists were not entirely unaware of the reinsertion of the MRIs used for the test-retest sample, which may have affected their performance during the second review. However, the neuroradiologists would not have known which MRIs were new and which were quality control studies given the blinding procedure used. Finally, the design of the SIT trial and the neuroradiology adjudication process did not allow us to examine specific factors that might have affected individual grader test-retest reliability. Specifically, we did not examine the relationship between lesion characteristics and inter- or intra-grader agreement because all three neuroradiologists did not individually record lesion number or perform volumetric measurements. For this reason, the reproducibility of detecting smaller, single lesions, which was associated with lower concordance on consensus ratings, is not known for individual graders.

In summary, inter- and intra-grader agreement among neuroradiologists is substantial and sufficient for detecting SCI in children with SCA. Our experience may not necessarily be extrapolated to other centers given the expertise of our neuroradiologists and the scientific rigor with which screening occurred in the SIT Trial. Nonetheless, these results provide support for using MRI as a reliable tool in screening children with SCA for SCI in the research setting and have important implications for the interpretation of results. Proper training of neuroradiologists and standardization of imaging protocols and operational definitions are necessary to ensure accuracy and reliability in identifying SCI by MRI, especially in the clinical setting. By adopting this approach, the use of MRI can and should be considered to improve detection of SCI, a prevalent morbid and progressive condition in this population.[18]

References

1. Ohene-Frempong K, Weiner SJ, Sleeper LA, et al. Cerebrovascular accidents in sickle cell disease: rates and risk factors. Blood. 1998; 91(1):288–294. [PubMed: 9414296]

2. Moser FG, Miller ST, Bello JA, et al. The spectrum of brain MR abnormalities in sickle-cell disease: a report from the Cooperative Study of Sickle Cell Disease. AJNR Am J Neuroradiol. 1996; 17(5):965–972. [PubMed: 8733975]

3. Pegelow CH, Macklin EA, Moser FG, et al. Longitudinal changes in brain magnetic resonance imaging findings in children with sickle cell disease. Blood. 2002; 99(8):3014–3018. [PubMed: 11929794]

4. Kwiatkowski JL, Zimmerman RA, Pollock AN, et al. Silent infarcts in young children with sickle cell disease. Br J Haematol. 2009; 146(3):300–305. [PubMed: 19500105]

5. Miller ST, Macklin EA, Pegelow CH, et al. Silent infarction as a risk factor for overt stroke in children with sickle cell anemia: a report from the Cooperative Study of Sickle Cell Disease. J Pediatr. 2001; 139(3):385–390. [PubMed: 11562618]

6. Armstrong FD, Thompson RJ Jr, Wang W, et al. Cognitive functioning and brain magnetic resonance imaging in children with sickle Cell disease. Neuropsychology Committee of the Cooperative Study of Sickle Cell Disease. Pediatrics. 1996; 97(6 Pt 1):864–870. [PubMed: 8657528]

7. White DA, Moinuddin A, McKinstry RC, Noetzel M, Armstrong M, DeBaun M. Cognitive screening for silent cerebral infarction in children with sickle cell disease. J Pediatr Hematol Oncol. 2006; 28(3):166–169. [PubMed: 16679942]

8. Schatz J, Brown RT, Pascual JM, Hsu L, DeBaun MR. Poor school and cognitive functioning with silent cerebral infarcts and sickle cell disease. Neurology. 2001; 56(8):1109–1111. [PubMed: 11320190]

9. Vendt BA, McKinstry RC, Ball WS, et al. Silent Cerebral Infarct Transfusion (SIT) trial imaging core: application of novel imaging information technology for rapid and central review of MRI of the brain. J Digit Imaging. 2009; 22(3):326–343. [PubMed: 18398653]

10. Vermeer SE, Longstreth WT Jr, Koudstaal PJ. Silent brain infarcts: a systematic review. Lancet Neurol. 2007; 6(7):611–619. [PubMed: 17582361]

11. Glauser TA, Siegel MJ, Lee BC, DeBaun MR. Accuracy of neurologic examination and history in detecting evidence of MRI-diagnosed cerebral infarctions in children with sickle cell hemoglobinopathy. J Child Neurol. 1995; 10(2):88–92. [PubMed: 7782614]

12. De Schryver EL, van Gijn J, Kappelle LJ, Koudstaal PJ, Algra A. Severity of cerebral white matter lesions and infarcts in patients with transient or moderately disabling cerebral ischaemia: reproducibility of grading by neurologists. Eur J Neurol. 2006; 13(8):901–903. [PubMed: 16879303]

13. Kapeller P, Barber R, Vermeulen RJ, et al. Visual rating of age-related white matter changes on magnetic resonance imaging: scale comparison, interrater agreement, and correlations with quantitative measurements. Stroke. 2003; 34(2):441–445. [PubMed: 12574557]

14. Wardlaw JM, Ferguson KJ, Graham C. White matter hyperintensities and rating scales-observer reliability varies with lesion load. J Neurol. 2004; 251(5):584–590. [PubMed: 15164192]

15. Ay H, Arsava EM, Vangel M, et al. Interexaminer difference in infarct volume measurements on MRI: a source of variance in stroke research. Stroke. 2008; 39(4):1171–1176. [PubMed: 18292377]

16. Neumann AB, Jonsdottir KY, Mouridsen K, et al. Interrater agreement for final infarct MRI lesion delineation. Stroke. 2009; 40(12):3768–3771. [PubMed: 19797188]

17. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33(1):159–174. [PubMed: 843571]

18. Debaun MR, Armstrong FD, McKinstry RC, Ware RE, Vichinsky E, Kirkham FJ. Silent cerebral infarcts: a review on a prevalent and progressive cause of neurological injury in sickle cell anemia. Blood. 2012; 119(20):4587–4596. [PubMed: 22354000]

19. Sasaki M, Hirai T, Taoka T, et al. Discriminating between silent cerebral infarction and deep white matter hyperintensity using combinations of three types of magnetic resonance images: a multicenter observer performance study. Neuroradiology. 2008; 50(9):753–758. [PubMed: 18551287]
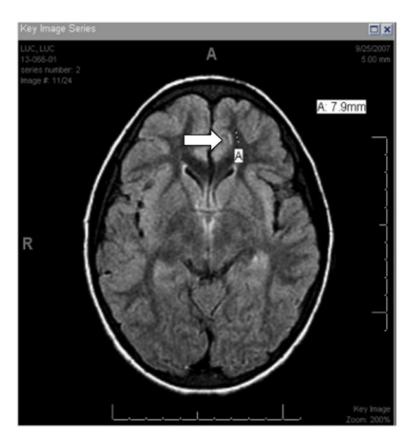
Figure 1.
Consensus decision for negative MRI. Example of MRI in which one grader voted for a qualifying lesion (A, arrow) in the subcortical region of the left frontal lobe but other two graders did not see the lesion. Consensus decision reached was MRI contained no silent cerebral infarct.
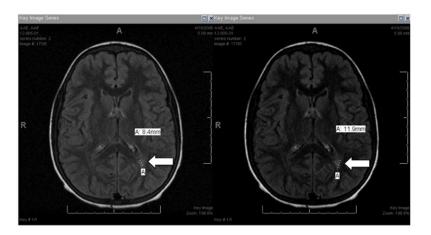
Figure 2.
Consensus decision for positive MRI. Separate views of an MRI in which two graders detected a qualifying lesion (A, arrow) in the periventricular region of the left parietal lobe but one grader did not. Consensus decision reached was MRI contained a silent cerebral infarct.

Table 1

Demographic Characteristics of Screening (N=1073), Test-Retest (N=53) and Discordant (N=8) Samples

Variable	Screening Sample	Test-Retest Sample	Discordant Sample[b]
Age (years)	9.1 ± 2.5[a]	9.0 ± 2.4	10.1 ± 2.3
Male gender (%)	546 (50.9)	29 (54.7)	4 (50.0)
Race (%)			
Black or African American	1017 (94.8)	49 (92.5)	8 (100)
Asian	6 (0.6)	1 (1.9)	0 (0)
Native Hawaiian/Other Pacific Islander	1 (0.1)	0 (0)	0 (0)
White	6 (0.6)	0 (0)	0 (0)
Other	43 (4.0)	3 (5.7)	0 (0)
Ethnicity (%)			
Hispanic or Latino	14 (1.3)	0 (0)	0 (0)
Not Hispanic or Latino	1056 (98.4)	53 (100)	8 (100)
Unknown	3 (0.3)	0 (0)	0 (0)
Scan type			
1.5Tesla	1031 (96.1)	51 (96.2)	8 (100)
3.0Tesla	42 (3.9)	2 (3.8)	0 (0)

[a]Standard Deviation,

[b]Discordant results for test-retest sample between first and second consensus ratings

Table 2

Intra-grader Agreement[a] of Test-Retest Sample (N=53)

		Second review			
		Yes	No	Indeterminate	Total	
Grader 1					
Yes		17	2	0	19	κ value[b] = 0.57
No		9	23	0	32	95% CI (0.35, 0.78)
Indeterminate		0	1	1	2	
Total		26	26	1	53	
Grader 2		Yes	No	Indeterminate	Total	
Yes		21	1	0	22	κ value = 0.76
No		6	25	0	31	95% CI (0.56, 0.92)
Indeterminate		0	0	0	0	
First Review Total		27	26	0	53	
Grader 3		Yes	No	Indeterminate	Total	
Yes		20	1	0	21	κ value = 0.74
No		5	25	0	30	95% CI (0.56, 0.91)
Indeterminate		1	1	0	2	
Total		26	27	0	53	
Consensus		Yes	No	Total		
Yes		21	2	23		κ value = 0.70
No		6	24	30		95% CI (0.51, 0.90)
Total		27	26	53		

[a] Agreement testing between first and second review of test-retest MRI sample,

[b] Calculated kappa values are weighted for graders and simple for consensus agreement testing

Table 3

Agreement[a] Testing between Grader and Consensus Rating by MRI Assessment Year (N=1073)

Year	N=	Grader 1		Grader 2		Grader 3	
		κ	95% CI	κ	95% CI	κ	95% CI
2004	6	1.00	1.00, 1.00	0.83	0.54, 1.00	1.00	1.00, 1.00
2005	271	0.86	0.79, 0.92	0.91	0.86, 0.96	0.85	0.79, 0.91
2006	268	0.90	0.84, 0.95	0.85	0.79, 0.92	0.93	0.89, 0.97
2007	219	0.85	0.78, 0.92	0.85	0.78, 0.92	0.91	0.85, 0.96
2008	150	0.88	0.81, 0.96	0.90	0.83, 0.97	0.91	0.84, 0.97
2009	137	0.90	0.83, 0.98	0.82	0.72, 0.92	0.94	0.89, 1.00
2010	22	0.95	0.87, 1.00	0.82	0.58, 1.00	0.96	0.87, 1.00
Overall	1073	0.88	0.85, 0.91	0.87	0.84, 0.90	0.91	0.88, 0.93

[a]Calculated kappa values are weighted