Title

Evaluating hippocampal internal architecture on MRI: Inter-rater reliability of a proposed scoring system

Journal

Epilepsy Research, 106(1-2)

ISSN

0896-6974

Authors

Hoef, Lawrence W Ver
Paige, A LeBron
Riley, Kristen O
et al.

Publication Date

2013-09-01

DOI

10.1016/j.eplepsyres.2013.05.009

Peer reviewed

Published in final edited form as:
Epilepsy Res. 2013 September ; 106(0): 146–154. doi:10.1016/j.eplepsyres.2013.05.009.

Evaluating hippocampal internal architecture on MRI: inter-rater reliability of a proposed scoring system

Lawrence W. Ver Hoef, MD[1,2], **A. LeBron Paige, MD**[3], **Kristen O. Riley, MD**[4], **Joel Cure, MD**[5], **Mehdi Soltani, MD**, **Frank B. Williams**[6], **Richard E. Kennedy, MD, PhD**[6], **Jerzy P. Szaflarski, MD, PhD**[1], and **Robert C. Knowlton, MD, MSPH**[7]

Lawrence W. Ver Hoef: LVERHOEF@UAB.EDU; A. LeBron Paige: LeBron-Paige@uiowa.edu; Kristen O. Riley: KRILEY@UAB.EDU; Joel Cure: CUREJ@UAB.EDU; Mehdi Soltani: soltanim_20@yahoo.com; Frank B. Williams: FBW@UAB.EDU; Richard E. Kennedy: RKennedy@ms.soph.uab.edu; Jerzy P. Szaflarski: SZAFLAJ@UAB.EDU; Robert C. Knowlton: Robert.C.Knowlton@uth.tmc.edu

[1]UAB Epilepsy Center, Department of Neurology, University of Alabama at Birmingham, Birmingham, AL

[2]Neurology Service, Birmingham VA Medical Center, Birmingham, AL

[3]Iowa Comprehensive Epilepsy Program, Department of Neurology, University of Iowa, Iowa City, IA

[4]Division of Neurosurgery, University of Alabama at Birmingham, Birmingham, AL

[5]Department of Radiology, University of Alabama at Birmingham, Birmingham, AL

[6]School of Medicine, University of Alabama at Birmingham, Birmingham, AL

[7]Texas Comprehensive Epilepsy Program, Department of Neurology, University of Texas at Houston, Houston, TX

Abstract

Background—Asymmetry of hippocampal internal architecture (HIA) has been reported to be a frequent imaging finding in epilepsy patients with temporal lobe epilepsy (TLE) who exhibit other signs of hippocampal sclerosis. HIA asymmetry may also be an independent predictor of the side of seizure onset in patients with otherwise normal MRI scans. The study of HIA asymmetry and its relationship to the laterality of TLE would benefit from a reliable method of assessing the clarity of HIA in MRI scans. We propose a visual scoring system that rates HIA clarity from 1 (imperceptible) to 4 (excellent) and report the inter-rater reliability (IRR) of this system.

Methods—In the initial preliminary phase of this study we examined IRR using a kappa statistic (κ) among a mixed group of expert and non-expert reviewers using only a brief description of the scoring system to score single images from a series of patients. In the second phase we explored the effect of training on the use of our HIA scoring system by assessing IRR among neuroimaging experts before and after a brief interactive training session. In this phase, multiple slices from each patient were scored. Separate κ values and intraclass correlation coefficients (ICC) were calculated from the scores given to each hippocampal image and from the asymmetry of scores

Corresponding Author: Lawrence W. Ver Hoef, MD, phone: 1-205-934-3866, fax: 1-205-975-6255, LVERHOEF@UAB.EDU, Address: Suite 312 CIRC, 1719 6th Avenue South, Birmingham, AL, 35205.

between left and right for each slice. In the third phase the effect of training on non-expert reviewers was explored using a similar approach as with the expert reviewers.

Results—In the preliminary phase of the study, HIA scoring of single images showed substantial agreement among expert reviewers (κ_{HIA}=0.65), fair agreement among non-expert reviewers (κ_{HIA}=0.27), and a fair to moderate degree of agreement among all the reviewers as a whole (κ_{HIA}=0.40). In the second phase, prior to training there was substantial agreement among expert reviewers in regard to the individual HIA scores (κ_{HIA}=0.62; ICC_{HIA}=0.81) but only moderate agreement on the degree of asymmetry (κ_{Asym}=0.47; ICC_{Asym}=0.71). Training improved agreement on the individual HIA scores (κ_{HIA}=0.58–0.72; ICC_{HIA}=0.76–0.84) and on the degree of asymmetry (κ_{Asym}=0.61–0.67; ICC_{Asym}=0.81–0.85). Among non-expert reviewers, scores improved from only a fair degree of agreement pre-training (κ_{HIA}=0.25, κ_{Asym}=0.25; ICC_{HIA}=0.68, ICC_{Asym}=0.66) to a moderate level of agreement after training (κ_{HIA}=0.54, κ_{Asym}=0.52; ICC_{HIA}=0.78, ICC_{Asym}=0.81).

Conclusions—The proposed HIA scoring system has a substantial degree of inter-rater reliability among experienced neuroimaging reviewers. Training improves the detection of asymmetries in HIA score in particular. Non-expert reviewers can employ the system with a moderate degree of reliability, and training has an even greater impact on the improvement of scoring reliability.

1. Introduction

Temporal lobe epilepsy (TLE) is the most common type of localization-related epilepsy in adults. Evidence of hippocampal sclerosis (HS) on MRI is a frequent and important finding in TLE. It indicates a high likelihood of having medically refractory epilepsy [1] but also a high likelihood of seizure freedom if the patient undergoes temporal lobectomy [2]. Strictly speaking, HS is a pathologic diagnosis based on microscopic tissue examination [3], but hippocampal atrophy and/or T2 signal hyperintensity on MRI are considered markers for HS [4]. It has been suggested that asymmetric loss of hippocampal internal architecture (HIA) clarity may be a third MRI hallmark of HS [4, 5].

HIA in this case refers to the laminar appearance of gray and white matter in coronal sections through the body of the hippocampus that arises from the structure of Ammon's horn. Specifically, the white matter tracts of the strata radiatum, lacunosum and moleculare on the inner surface of the subiculum and CA1-CA3 in apposition with the stratum moleculare of the dentate gyrus create a hypointense (dark) band on T2-weighted images [6, 7] that form a typical spiral appearance when seen in its entirety (Figure 1). This hypointense band is commonly on the order of 1 mm or less in thickness in coronal section and therefore is on the margin of what can be easily resolved with conventional MRI sequences. As such, it is not uncommon that HIA is not clearly visible on MR images of normal hippocampi. Even images acquired at high field strength (3T) with sub-millimeter resolution may fail to demonstrate any HIA in some individuals (Figure 2a), while HIA may be quite clear in other individuals (Figure 2b). This variability in HIA clarity may even be observed between adjacent slices in a single individual, even in those with no significant difference in clarity between the left and right side within a given slice (Figure 3). The factors that account for this symmetric variability are not known, but it is most likely related to individual anatomic variations in thickness of the hippocampal sublayers, MR tissue characteristics, and volume averaging effects that arise from through-plane variation across the relatively thick slices required for high-resolution imaging.

Of particular interest to epilepsy and neuroimaging specialists is whether or not an asymmetry of HIA clarity in the absence of hippocampal atrophy or signal abnormality has significance in identifying pathology related to TLE. In order to assess the significance of an

HIA asymmetry, a reliable measure of HIA clarity is necessary to quantify the degree of asymmetry and statistically analyze its relationship to the laterality of seizure onset. In this work we propose a 4-point visual scoring system that rates HIA clarity from 1 (imperceptible) to 4 (excellent). We report the inter-rater reliability (IRR) of this system when applied by both expert and non-expert reviewers, and the impact of a brief training session on reliability. The predictive value of HIA asymmetry in TLE will be described in a separate related report [8].

2. Methods

After Institutional Review Board approval was obtained, the assessment of inter-rater reliability (IRR) was conducted in three phases. The initial phase consisted of a preliminary assessment of IRR between five reviewers with various backgrounds related to epilepsy neuroimaging. Two reviewers (LV and RCK) are experienced epileptologists with additional qualification in neuroimaging (American Society of Neuroimaging) who routinely read and interpret clinical epilepsy MRI scans as the physician of record. These reviewers will be referred to as “expert” reviewers. The other three reviewers included an epileptologist (ALP) and an epilepsy neurosurgeon (KOR), each with more than five years of post-fellowship experience including routine clinical review of MRI scans of TLE patients but not responsibility for official MRI interpretation, and an epilepsy fellow who was a board-eligible neurologist (MS). These three reviewers will be referred to as “non-expert” reviewers.

The images for review were taken from a high-resolution T2-weighted TSE sequence (TR 3000/ TE 110/ flip angle 90/NEX 2/FOV 240 mm/acquisition matrix 912×912/ reconstruction matrix 1024×1024/ slice 3 mm/gap 1 mm). This sequence was chosen for its high in-plane resolution of 0.26 mm. Slices were oriented in an oblique coronal plane orthogonal to the long axis of the hippocampus for optimal imaging the body of the hippocampus in cross-section. Images were obtained retrospectively from clinical scans acquired with our institution’s temporal lobe MRI protocol on a Philips Achieva 3T scanner.

For the initial phase of this study, scans from 20 consecutive patients that were free of significant movement artifact or significant temporal lobe pathologic findings were used. These patients were referred from the general neurology clinic, epilepsy clinic, or other neurology sub-specialty clinics, and because the purpose of this study was simply to assess the reliability of the scoring system and not to correlate HIA scores with the presence of epilepsy, patients were included without regard to the presence or absence of a definitive or suspected diagnosis of epilepsy. A single slice through the body of both hippocampi that best represented the HIA clarity for each patient’s hippocampi was selected for scoring. Preselection of a single slice, as opposed to using all available slices, potentially introduces bias, but this was felt to be acceptable in this phase of the study which was simply a preliminary exploratory investigation intended to determine if at least a moderate degree of inter-rater agreement was present and worth further investigation. The results of this phase would influence the design of the subsequent phases of investigation.

Reviewers were given a written description of the HIA scoring system and a single example image of each level of HIA clarity as shown in Figure 4. Reviewers independently scored each hippocampus in each image and were blinded to any clinical information about the patients from whom the images were acquired. The scores were tabulated and a Fleiss’ kappa statistic [9] was calculated for the expert group, the non-expert group, and all reviewers as a group. All references to kappa statistics in this study refer to a Fleiss’ kappa statistic as well.

The purpose of the second phase of the study was to conduct a more thorough and detailed assessment of IRR and determine if experience and training improved the IRR among expert reviewers. A second goal was to establish a set of multiple example images of each of the four levels of the scoring system to be used as a reference in applying the scoring system. Three expert reviewers were used in this phase, two of which were the expert reviewers from the first phase plus a third reviewer was added who is a board certified neuroradiologist (JC). Images for this phase of the study were taken from 24 consecutive clinical scans performed on patients with suspected or confirmed TLE that were free of significant movement artifact. In this phase three consecutive slices through the body of the hippocampus from each patient were reviewed instead of a single representative slice as was used in the first phase. Reviewing multiple slices from each patient more closely reflects clinical practice and reduces the bias introduced by selecting only a single representative slice as was done in the previous preliminary phase of the study. Three separate image sets with eight patients in each set (3 slices per patient, 24 total slices per set) were used for this phase. These will be referred to as sets A, B, and C respectively.

Image set A was independently viewed and scored by the expert reviewers based on the same written description and single examples of the scoring system used in the preliminary phase (Figure 4) and the scores were recorded. The reviewers then viewed each of the images in set B while sitting together viewing the images on the same computer monitor. As each image was viewed, the reviewers stated how they would score each hippocampal image. Of note, the patients that comprised image set B were different patients than those previously reviewed in set A, and the images and scores from set A were deliberately not discussed so as to avoid biasing the reviewers' subsequent post-training scoring of set A. Given that there is some degree of subjectivity in the application of a visual rating system, it was expected that there would be some differences of opinion. When disagreement on scores occurred, the reasons for scoring one way versus another were discussed among the reviewers until unanimous consensus was reached. All 24 slices (48 individual hippocampal images) were reviewed, discussed, and rated in a single 90-minute session. There were no cases where consensus was not reached within a few minutes of discussion nor was there a single reviewer who had a dissenting opinion more frequently than any other. The primary purpose of this was for the reviewers to work out how they apply the rules of the rating system and to fine-tune their judgment in light of the converging opinions of the other reviewers. The secondary purpose was to develop a reference image set containing multiple examples of each level of the scoring system that would hopefully cover the spectrum of HIA clarity within each level. To this end, a box was drawn around each hippocampus -- one on the left and another on the right -- and "cut" out of each image and "pasted" into a composite image of all of the hippocampal slices organized by the consensus score (Appendix). This reference image set would then provide benchmark images to help in applying ratings consistently henceforth.

After the training session, the images of set A were randomly rearranged in order, and half of the images were randomly selected to be flipped left to right. This newly randomized image set A was then scored again by the expert reviewers separately. The rearrangement of order and orientation of the images was intended to reduce the potential for bias from a reviewer recognizing the order or appearance of the images from the pre-training scoring session. This approach was used instead of simply using a completely different image set to avoid the possibility that the images in the second set were intrinsically more or less challenging to score than the first, which would skew the post-training results. The reviewers were also instructed to have the reference composite image displayed on a second monitor and to refer to this standard as needed during the post-training scoring session.

In addition to the HIA scores rendered for each hippocampus in a given slice, an HIA asymmetry score was calculated for each slice by subtracting the right HIA score from the left. Using this method a positive HIA asymmetry score indicates a loss of HIA clarity on the left and a negative HIA asymmetry score indicates a loss of HIA clarity on the right.

In the second phase, Kappa statistics and intraclass correlation coefficients (ICC(2,1) as described by Shrout and Fleiss [10]) were calculated for both the "raw" HIA scores (κ_{HIA}, ICC_{HIA}; N=48) and the HIA asymmetry score (κ_{Asym}, ICC_{Asym}; N=24). This was done first for the pre-training scoring of image set A and then for the post-training scoring of image set A with all of the images in random order. Unexpectedly, the κ_{HIA} and ICC_{HIA} were lower in the post-training review than in the pre-training review. We suspected that this decline was most likely due to the fact that the post-training images were reviewed in a completely random order such that the reviewers did not have the benefit of seeing the adjacent slices from the same patient as they did in the pre-training review. Therefore we repeated the post-training review with randomization limited to the *order of patients* while keeping each patient's three slices grouped in the original anatomical order as shown in Table 1. To confirm that any improvement was not related to increased familiarity with the images in set A, the reviewers scored a third set of images (set C) to confirm the findings.

In the third phase of the study we sought to determine the impact of training and availability of the composite reference image on the IRR among non-expert reviewers. The same three non-expert reviewers from the first phase independently scored the images of set C and these scores were recorded. The non-expert reviewers then met together with one of the expert reviewers (LV) for a 30-minute training session to discuss their impressions of the scoring system, review the reference image set, and go over images from three example patients (three slices per patient, nine images total) from set A. They then did a post-training review and scoring of image set C with the patients in random order and half of the patients' images were flipped right to left, but all three images for each patient were again kept in original anatomic order as illustrated in the third column of Table 1. As in the second phase, kappa statistics and ICCs were calculated on the pre- and post-training scores for both HIA scores and asymmetry scores.

The kappa statistic indicates the degree to which the observed agreement between reviewers is greater than would be expected to occur by chance, with $\kappa = 1.0$ indicating perfect agreement and $\kappa = 0.0$ indicating no greater than random chance. There is no universally accepted cut-off value for kappa statistics that indicate an acceptable level of agreement [11]. The most commonly used ranges are those proposed by Landis and Koch [12], which are 0.0–0.20 as slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.0 as almost perfect agreement. Similarly, there is no widely agreed upon range of acceptable values for ICC. Some have suggested that an ICC of 0.4 to 0.75 is "fair to good", and >0.75 is "excellent" [13] and recommend values of 0.75 or greater for health research [14], while yet others use the following ranges: 0.5–0.6 indicates moderate agreement; 0.7–0.8 indicates strong agreement; and >0.8 indicates almost perfect agreement [12]. When two kappa statistics or ICCs are reported, a measure of statistical significance of the difference (e.g. p-value) is rarely given and there is some controversy about how to use and compare values as measures of level of agreement [15, 16]. Confidence intervals narrow enough to show interval improvements require extremely large sample sizes [11]. As such, significance of differences between measures of agreement are not reported here. This work is intended to be exploratory and descriptive, and a larger prospective study would more accurately define the level of agreement between reviewers.

3. Results

In the initial phase of the study the kappa statistic was calculated between two expert reviewers, three non-expert reviewers, and between all five of the reviewers together. This showed that there was substantial agreement among expert reviewers (κ_{HIA} = 0.65), fair agreement among non-expert reviewers (κ_{HIA} = 0.27), and a fair to moderate degree of agreement among all the reviewers as a whole (κ_{HIA} = 0.40).

The difference in agreement between expert and non-expert reviewers in the first phase supported a second phase investigation into the effect of experience and training on IRR. Demographic and clinical information for patients whose images were used in the second and third phase from which the conclusions of this study are drawn are listed in Table 2. The results of the second of the study are summarized in Table 3, along with the results of the third phase. There was substantial agreement among expert reviewers in the pre-training use of the HIA scoring system on individual hippocampal images in image set A (κ_{HIA} = 0.62; ICC_{HIA}=0.81). However, the level of agreement was not as high for HIA asymmetry among expert reviewers before training (κ_{Asym} = 0.47; ICC_{Asym}=0.71). In the initial post-training scoring of image set A, in which all of the images were in random order, the level of agreement on the HIA score unexpectedly decreased (κ_{HIA} = 0.46; ICC_{HIA}=0.72) and the agreement on HIA asymmetry also did not improve (κ_{Asym} = 0.47; ICC_{Asym}=0.65). In the revised post-training review of image set A, in which only the order of the patients was random but any given patient's images were shown in original sequence, the agreement on HIA scores improved from the initial post-training review and was similar to that of the pre-training review (κ_{HIA} = 0.58; ICC_{HIA}=0.76), and the agreement on the degree of asymmetry improved substantially (κ_{Asym} = 0.61; ICC_{Asym}=0.81). The last portion of this phase was a review of image set C, which was previously unseen, and showed slightly higher agreement for both HIA scores (κ_{HIA} = 0.72; ICC_{HIA}=0.84) and asymmetry scores (κ_{Asym} = 0.67; ICC_{Asym}=0.85).

In the third phase of the study we examined the effect of training and availability of the HIA reference image set on the IRR of non-expert reviewers. The pre-training review of image set C showed a lower degree of agreement between reviewers as compared to the pre-training review among experts for both the HIA score (κ_{HIA} = 0.25; ICC_{HIA}=0.68) and the asymmetry score (κ_{Asym} = 0.25; ICC_{Asym}=0.66). After training both HIA scores and asymmetry scores for non-expert reviewers improved markedly (κ_{HIA} = 0.54, ICC_{HIA}=0.78; κ_{Asym} = 0.52, ICC_{Asym}=0.81).

4. Discussion

The main goal of this study was to develop and demonstrate the reliability of a scoring system for measuring the degree of hippocampal internal architecture asymmetry. In the preliminary phase we intended to gain a basic understanding of how easy the system is to use and determine if there is an obvious difference in scoring between reviewers of different backgrounds. This phase showed that there was substantial agreement among reviewers who are neuroimaging experts, but relatively poor agreement among the non-expert reviewers. However, except for the first author, the reviewers in this portion of the study had no prior experience in using the HIA scoring system and were given no training. Rather they received only a simple description of the scale and a single example of each level (Figure 4), which raised the question whether the IRR would be greater for both expert and non-expert reviewers if they had more training and experience using the HIA scoring system.

In the second phase of the study we examined the effect of experience and training on the IRR of the HIA scoring system specifically among expert reviewers. We also changed the composition of the sets of reviewed images to include three consecutive slices from each

subject instead of a single representative slice from each subject. This change was made to more closely emulate clinical practice and to avoid the possibility of artificially raising the IRR by selecting only the easiest image to interpret from each patient for scoring. In the pre-training review, there was substantial agreement on the scoring of HIA among experts, which was similar to what was seen in the preliminary phase, but there was only moderate agreement on the degree of asymmetry before training. Surprisingly, the first post-training scoring demonstrated *less* agreement than the pre-training scoring.

The reviewers considered possible sources of bias and felt that the most likely cause of the decrease in agreement was that the images in the post-training review were in completely random order, as opposed to the pre-training review in which consecutive slices from a given subject were reviewed in sequential order. We hypothesized that as reviewers scored each image they were consciously or unconsciously using information from the previous slice(s) of the same patient to make their judgments regarding the HIA score and degree of asymmetry. When the images were reviewed keeping each subject's images in sequential order and only randomizing the order of the subjects, the agreement in HIA scoring of image set A improved from the initial post-training level, but only to a level similar to what was seen before training. However, the agreement on the degree of asymmetry in the revised post-training review did improve noticeably from the pre-training level. The use of adjacent slices to interpret a given slice should not be considered "cheating" as it reflects the way images are interpreted in clinical practice -- when a subtle or questionable MRI finding is seen on more than one consecutive slice it is given more credence than if it is seen in one slice alone.

Because image set A was reviewed a total of three times, we felt it was necessary to confirm these findings with the review of a new set of images (image set C) that had not been previously reviewed. This produced somewhat higher kappa values for both the HIA score and the asymmetry score, confirming the capacity of experts to have substantial agreement using this scoring scheme.

The pre-training review by the non-expert reviewers resulted in only a fair level of agreement, and in four of the scores the reviewers differed in their scoring by more than a single level. However, after only brief training and with use of the reference image set the agreement among non-experts showed substantial improvement into the moderate range.

Several observations can be made from this study. These findings suggest that among expert reviewers, even the limited amount of training that occurred in the study improved their ability to assess the asymmetry of HIA consistently, though there was not obvious improvement in the agreement of absolute HIA scores, which was already substantial. Among the non-expert reviewers the benefit of even minimal training and use of benchmark images in the reference image set produced a dramatic improvement in agreement in both HIA scores and asymmetry scores, which were rather low prior to training. Additional training would likely produce even greater improvements for both groups, though it is not clear how much the improvements, particularly for the non-experts, were related to training or to using the reference image set. The advantage of having a reference image set is that it shows the spectrum of images within each level of the scoring system and helps give boundaries for the intrinsically subjective application of the scoring system criteria. For future study and eventual clinical use the appendix of this report will be available on the internet and made widely accessible. This may serve as a guide for clinicians and researchers as they gain experience and hone their judgment in applying the scoring system, and help provide consistency in formal studies of the clinical significance of asymmetry of HIA.

In this study we calculated kappa statistics and ICCs for both the "raw" HIA score (the score of clarity of HIA in each hippocampal image) and the asymmetry score (the difference between the right and left HIA score). These values are related but they are not identical nor are they a simple linear transformation of each other. If there were perfect agreement in HIA scoring, there would be perfect agreement in the asymmetry scores, but when there is less than perfect agreement, there may be cases where reviewers agree on the degree of asymmetry but disagree on the raw HIA scores. For example, if there is a clear asymmetry between left and right sides but the left side is on the border between being scored as a "1" or a "2" and the right is on the border between being scored as a "2" or a "3", one reviewer could score the left as a "1" and the right as a "2", while another reviewer could score the left as a "2" and the right as a "3". This would result in a disagreement between reviewers on the HIA scores, but agreement on the degree of asymmetry. In this case one reviewer was judging both sides more conservatively and the other less conservatively, but both agree on the degree and laterality of asymmetry. It is likely that the human eye can better differentiate subtle asymmetries in HIA than it can categorize HIA clarity into a somewhat subjective 4-point system, and in fact the intended utility of this system is to assess HIA asymmetry as a marker of unilateral hippocampal pathology. In clinical practice, it is routine for the normality of structures seen on coronal or axial section to be evaluated in comparison to the appearance of the homologous structure on the contralateral side as an internal control. In light of this, the reliability of the HIA asymmetry score may arguably be the more important measure.

Of note, HIA is an ordinal variable, meaning that the categories have a distinct order and difference between them (the difference between a "1" and a "3" is greater than between a "1" and a "2"). When using a kappa statistic to assess agreement with an ordinal variable, one would ideally use a weighted kappa in which there is a greater penalty for a large difference in scores between reviewers than for a difference in scores of only one level. However, when more than two reviewers are used there is no way to calculate a weighted kappa and only an unweighted kappa can be used. As such, the unweighted kappa becomes a measure of *exact agreement* only and no credit is given for scores that are close, which is overly stringent if most or all of the disagreements are slight. Alternatively, an ICC takes into account scores that are close and does not invoke as much of a penalty if one reviewer is consistently more conservative or consistently more lenient in his/her ratings. In our case, this resulted in ICC values that are higher than the kappa values, but with less of a spread between them. We chose to report both measures because the ICCs show that there is a very strong correlation between raters' scores (all post-training ICCs were >0.75), but the more strict criterion of exact agreement with the unweighted kappa more clearly illustrates differences between the pre- and post- states and between expert and non-expert reviewers. It is also worth mentioning that the trends in changes from testing state to testing state are quite similar between the ICCs and the kappas even though the values are different.

The main limitation of this study is the subjective nature of the scoring system. The descriptors we have chosen to define each level are admittedly "artificial" in the sense that they do not follow from an obvious biological parameter. These descriptors are simply an attempt to mark waypoints along the spectrum of HIA clarity to provide a metric by which HIA clarity can be assessed. Given that HIA clarity is a continuous spectrum, there are certainly cases that will fall on the borderline between our defined categories. Our purpose in showing a large number of examples in the appendix is not to show idealized textbook examples that fall clearly within the description of each level, but rather to give real world examples of images that were subjectively determined by consensus to be in the stated category. Though differences between levels of HIA clarity may be subtle, the fact that this rating system is shown to have good inter-rater reliability indicates that results are

reasonably reproducible and the method may be useful to study phenomena with moderate effect sizes.

Quantitative measures of evaluating the structure of the hippocampus exist, such as hippocampal subfield volumetry [17, 18], but these methods cannot be as easily employed as our system. Furthermore, most if not all studies of hippocampal subfield volumetry begin with a subjective step of manually tracing the hippocampal subfields and are therefore not entirely objective. We have included a brief review of other methods that have attempted to assess hippocampal internal structure in our report of a related study [8] and direct interested readers to the discussion section of that manuscript.

The T2-TSE sequence was chosen for this study because of its high in-plane resolution. MPRAGE and T1-inversion recovery sequences commonly used in epilepsy protocols often demonstrate HIA to some degree, but typically have an in-plane resolution of 1 mm (256 × 256) and 0.47 mm (512 × 512) respectively, while our T2-TSE is acquired at a resolution of 0.26 mm. The laminar bands that define HIA are often only 1 to 3 pixels in width even when acquired with a 0.26 mm resolution, particularly in the area near CA2-3, and may not be as well depicted with lower resolution scans. Determining the ideal sequence for imaging HIA would be a complicated process of balancing slice thickness, resolution, tissue contrast, signal-to-noise ratio, and scan time. The sequence we used is typical of sequences commonly available on 3T clinical scanners, which supports the general applicability of our findings, but a better sequence for demonstrating HIA may exist and could be an interesting area of future research.

5. Conclusions

The proposed HIA scoring system has a substantial degree of inter-rater reliability (κ_{HIA} = 0.58–0.72, κ_{Asym} = 0.61–0.67; ICC_{HIA}=0.76–0.84, ICC_{Asym}=0.81–0.85) among experienced neuroimaging reviewers with even a brief period of training in using the system. Training seems to improve the perception of asymmetries in HIA score in particular. Non-expert reviewers can also employ the system with a moderate degree of reliability, and training has a greater impact in the improvement in reliability of their scores. Viewing several sequential slices from a given subject, as is done in clinical practice, seems to improve reliability of the scores over viewing images in isolation. Additional work is necessary to more accurately define the reliability of this scoring system and to assess its value as a tool to detect clinically significant asymmetries of hippocampal structure.

Acknowledgments

Research reported in this publication was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under Award Number 1K23EB008452-01A1. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Abbreviations

HIA	hippocampal internal architecture
HS	hippocampal sclerosis
TLE	temporal lobe epilepsy
IRR	inter-rater reliability

References

1. Semah F, Picot MC, Adam C, Broglin D, Arzimanoglou A, Bazin B, Cavalcanti D, Baulac M. Is the underlying cause of epilepsy a major prognostic factor for recurrence? Neurology. 1998; 51:1256–62. [PubMed: 9818842]
2. Radhakrishnan K, So EL, Silbert PL, Jack CR Jr, Cascino GD, Sharbrough FW, O'Brien PC. Predictors of outcome of anterior temporal lobectomy for intractable epilepsy: a multivariate study. Neurology. 1998; 51:465–71. [PubMed: 9710020]
3. Margerison JH, Corsellis JA. Epilepsy and the temporal lobes. A clinical, electroencephalographic and neuropathological study of the brain in epilepsy, with particular reference to the temporal lobes. Brain. 1966; 89:499–530. [PubMed: 5922048]
4. Jackson GD, Berkovic SF, Duncan JS, Connelly A. Optimizing the diagnosis of hippocampal sclerosis using MR imaging. AJNR Am J Neuroradiol. 1993; 14:753–62. [PubMed: 8517369]
5. Jackson GD, Kuzniecky RI, Cascino GD. Hippocampal sclerosis without detectable hippocampal atrophy. Neurology. 1994; 44:42–6. [PubMed: 8290088]
6. Howe KL, Dimitri D, Heyn C, Kiehl TR, Mikulis D, Valiante T. Histologically confirmed hippocampal structural features revealed by 3T MR imaging: potential to increase diagnostic specificity of mesial temporal sclerosis. AJNR Am J Neuroradiol. 2010; 31:1682–9. [PubMed: 20538822]
7. Thomas BP, Welch EB, Niederhauser BD, Whetsell WO Jr, Anderson AW, Gore JC, Avison MJ, Creasy JL. High-resolution 7T MRI of the human hippocampus in vivo. J Magn Reson Imaging. 2008; 28:1266–72. [PubMed: 18972336]
8. Ver Hoef LW, Williams FBW, Kennedy R, Szaflarski JP, Knowlton RC. Predictive value of hippocampal internal architecture asymmetry in temporal lobe epilepsy. 2012 UNDER REVIEW.
9. Fleiss JL. Measuring nominal scale agreement among many raters. Psychological Bulletin. 1971; 76:378–382.
10. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979; 86:420–8. [PubMed: 18839484]
11. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther. 2005; 85:257–68. [PubMed: 15733050]
12. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33:159–74. [PubMed: 843571]
13. Fleiss, JL. Wiley classics library. Wiley classics library 1999. New York: Wiley; The design and analysis of clinical experiments; p. xivp. 432
14. Streiner, DL.; Norman, GR. Oxford medical publications. 2. Oxford; New York: Oxford University Press; 1995. Health measurement scales: a practical guide to their development and use; p. viiip. 231
15. Donner A, Zou G. Interval estimation for a difference between intraclass kappa statistics. Biometrics. 2002; 58:209–15. [PubMed: 11890316]
16. Uebersax, J. The Myth of Chance-Corrected Agreement. Kappa Coefficients 2009. Oct 01. 2009 [cited 2012 03 Oct]; Available from: http://www.john-uebersax.com/stat/kappa2.htm
17. Mueller SG, Laxer KD, Barakos J, Cheong I, Garcia P, Weiner MW. Subfield atrophy pattern in temporal lobe epilepsy with and without mesial sclerosis detected by high-resolution MRI at 4 Tesla: preliminary results. Epilepsia. 2009; 50:1474–83. [PubMed: 19400880]
18. Ekstrom AD, Bazih AJ, Suthana NA, Al-Hakim R, Ogura K, Zeineh M, Burggren AC, Bookheimer SY. Advances in high-resolution imaging and computational unfolding of the human hippocampus. Neuroimage. 2009; 47:42–9. [PubMed: 19303448]

Appendix

Reference image set showing individual hippocampal slices that were categorized by consensus of three neuroimaging experts into the four levels of the Hippocampal Internal Architecture Scoring System. Slices are scored from 4 (excellent) to 1 (imperceptible) based

on visual clarity of HIA. This provided benchmark examples of each level for post-training scoring, and may be used as a guide in applying the scoring system in the future.

Highlights

- We propose a 4-point scoring system for rating clarity of hippocampal internal architecture (HIA)
- The HIA scoring system has substantial inter-rater reliability among experienced reviewers
- The scoring system has moderate inter-rater reliability among non-expert reviewers
- Inter-rater reliability improved for both groups after a minimal amount of training

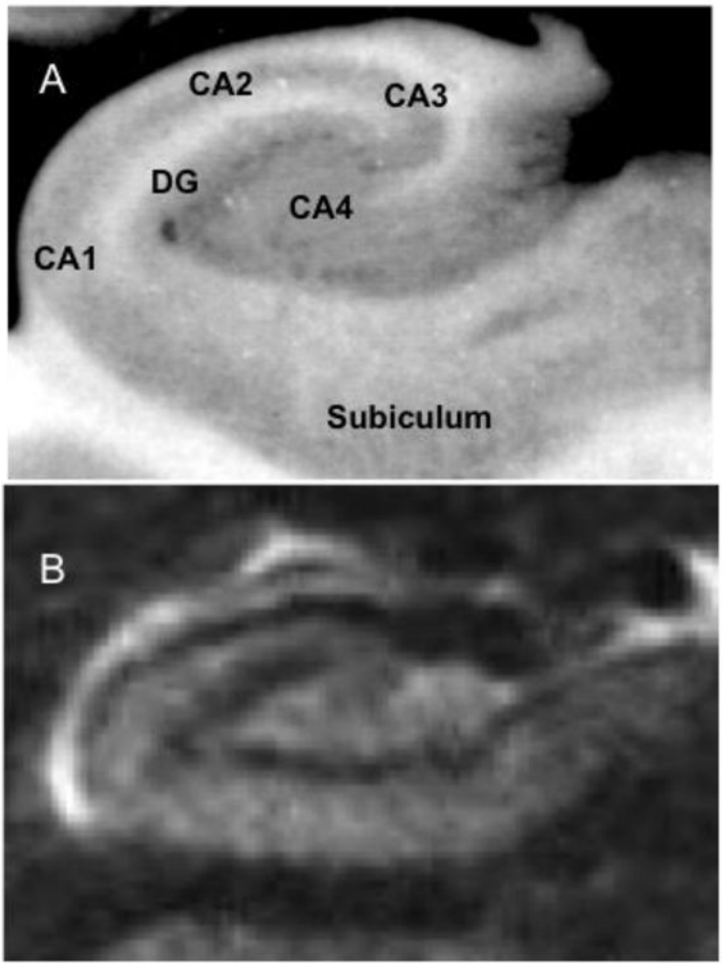

Figure 1.

A. Photograph of a coronal section through the right hippocampus of a cut brain showing the subfields of Ammon's horn. B. Magnified view of a high-resolution MRI coronal slice through the hippocampal body that shows clear differentiation of hippocampal internal architecture.

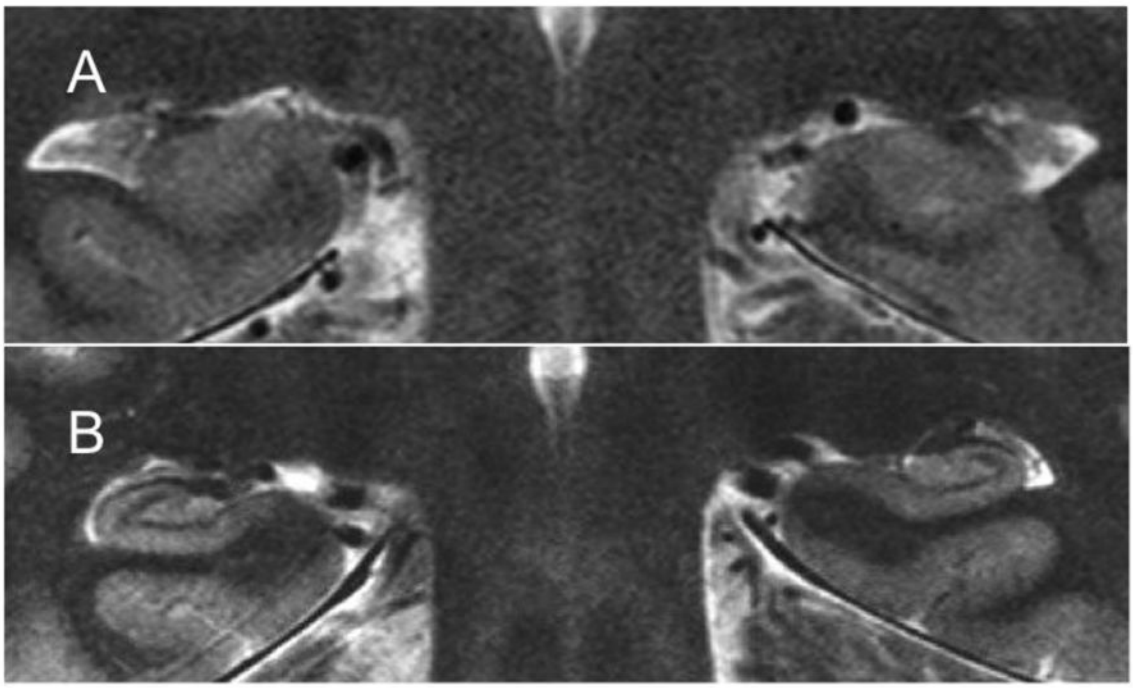

Figure 2.
T2-weighted coronal MR images of the hippocampi of two different individuals. The image shown in A has no perceptible HIA despite good image quality and no other hippocampal imaging abnormalities, while HIA is clearly defined in B.

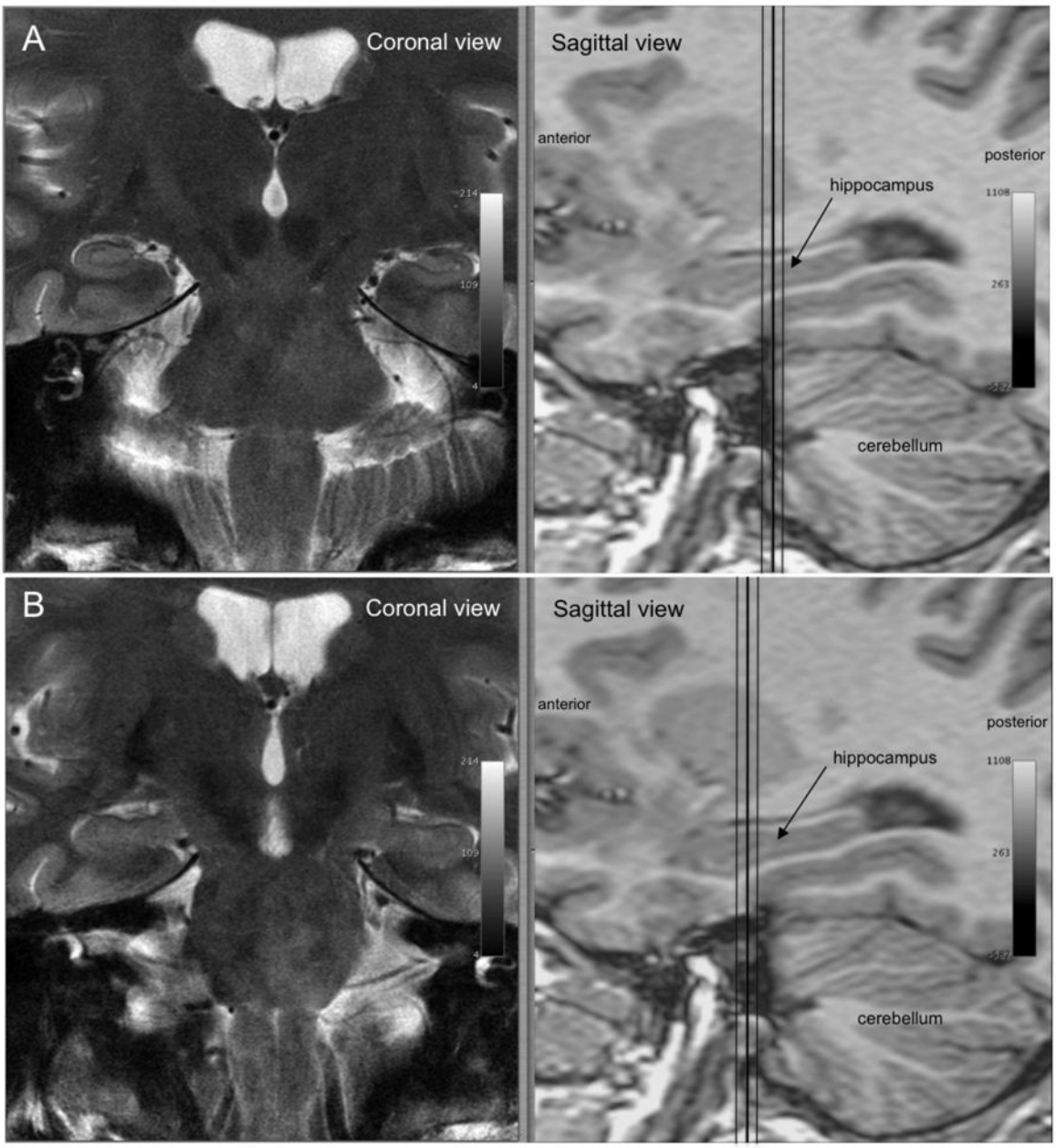

Figure 3.

Variability in differentiation of hippocampal internal architecture within the same scan. The upper left image (A) is a coronal slice showing *clear* differentiation of hippocampal internal architecture. The lower left image (B) is an adjacent slice from the same scan, which shows *poor* differentiation of hippocampal internal architecture. The images on the right are sagittal slices, each showing the location, orientation, and thickness (marked with the dark lines) of the coronal slice to its left. The A/P orientation and landmarks of hippocampus and cerebellum given for reference. Color bars indicate arbitrary units of image intensity. Images are shown in radiologic convention.

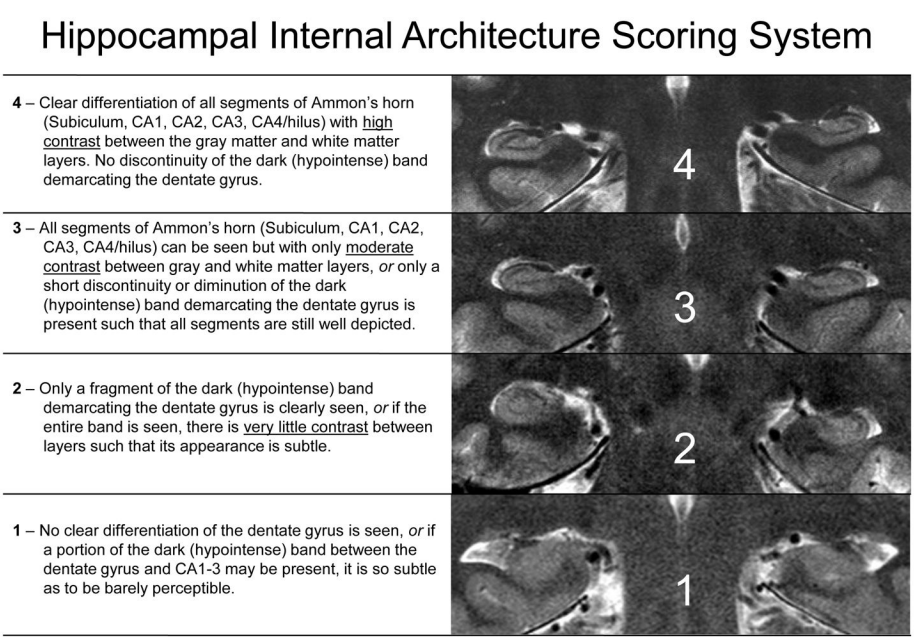

Figure 4.

Hippocampal Internal Architecture (HIA) scoring system. Left: A detailed explanation of each level; Right: Example of hippocampal images corresponding to each scoring level are presented with both sides in the example images having the same HIA scores. Images are presented in radiological convention.

Table 1

Order of viewing of images for each round of scoring. In the pre-training review, the three images from each subject are in order from anterior to posterior. In the initial post-training review, the images were in random order. In the revised post-training review, the subject order was random but each subjects images were in the original anterior-to-posterior order

	Pre-Training Review		Initial Post-Training Review		Revised Post-Training Review	
Order	Subject	Image	Subject	Image	Subject	Image
1	1	1	6	2	3	1
2		2	5	3		2
3		3	2	3		3
4	2	1	6	2	7	1
5		2	3	1		2
6		3	8	1		3
7	3	1	4	2	6	1
8		2	1	1		2
9		3	8	3		3
10	4	1	7	1	1	1
11		2	3	2		2
12		3	1	3		3
13	5	1	5	2	5	1
14		2	8	2		2
15		3	5	1		3
16	6	1	7	2	4	1
17		2	6	1		2
18		3	1	2		3
19	7	1	2	3	8	1
20		2	4	3		2
21		3	4	1		3
22	8	1	2	1	2	1
23		2	3	3		2
24		3	7	3		3

Table 2

Description of patients used in phase 2 and 3

Demographics	**N=24**
Male/Female	6/18
Age Range (Median)	18–58 (30.5)
Handedness	23 R, 1 L
Age at Onset (Median)	1–50 (25)
Duration of epilepsy (Median)	1–41 yrs (4.5 yrs)
Diagnosis	
Clinically suspected TLE, normal EEG	7 (29%)
Ictal EEG evidence of right TLE	5 (21%)
Ictal EEG evidence of left TLE	9 (38%)
Interictal EEG evidence of right TLE	0
Interictal EEG evidence of left TLE	3 (13%)
MRI findings	
Normal hippocampi (No atrophy or T2 signal hyperintensity)	18 (75%)
Hippocampal atrophy and T2 signal hyperintensity	3 (13%)
T2 signal hyperintensity in the end folium region but no atrophy	2 (8%)
Bilateral hippocampal malrotation	1 (4%)

Table 3

Effects of Training on inter-rater reliability of HIA scoring and HIA asymmetry scores

Expert Reviewers	K_{HIA}	K_{Asym}	ICC_{HIA}	ICC_{Asym}
Pre-training	0.62	0.47	0.81	0.71
Initial Post-training	0.46	0.47	0.72	0.65
Revised Post-training	0.58	0.61	0.76	0.81
Confirmation Set	0.72	0.67	0.84	0.85
Non-Expert Reviewers				
Pre-training	0.25	0.25	0.68	0.66
Post-training	0.54	0.52	0.78	0.81