Psychophysical Evaluation of Visual vs. Computer-Aided Detection of Brain Lesions on Magnetic Resonance Images

Chen Solomon, MSc,[1] Omer Shmueli, BSc,[1] Shai Shrot, MD,[2,3]
Tamar Blumenfeld-Katzir, PhD,[1] Dvir Radunsky, MSc,[1] Noam Omer, MSc,[1] Neta Stern, BSc,[1]
Dominique Ben-Ami Reichman, MD,[3] Chen Hoffmann, MD,[2,3] Moti Salti, PhD,[4,5]
Hayit Greenspan, PhD,[1] and Noam Ben-Eliezer, PhD[1,5,6,7]*

Background: Magnetic resonance imaging (MRI) diagnosis is usually performed by analyzing contrast-weighted images, where pathology is detected once it reached a certain visual threshold. Computer-aided diagnosis (CAD) has been proposed as a way for achieving higher sensitivity to early pathology.
Purpose: To compare conventional (i.e., visual) MRI assessment of artificially generated multiple sclerosis (MS) lesions in the brain's white matter to CAD based on a deep neural network.
Study Type: Prospective.
Population: A total of 25 neuroradiologists (15 males, age 39 ± 9, 9 ± 9.8 years of experience) independently assessed all synthetic lesions.
Field Strength/Sequence: A 3.0 T, T_2-weighted multi-echo spin-echo (MESE) sequence.
Assessment: MS lesions of varying severity levels were artificially generated in healthy volunteer MRI scans by manipulating T_2 values. Radiologists and a neural network were tasked with detecting these lesions in a series of 48 MR images. Sixteen images presented healthy anatomy and the rest contained a single lesion at eight increasing severity levels (6%, 9%, 12%, 15%, 18%, 21%, 25%, and 30% elevation in T_2). True positive (TP) rates, false positive (FP) rates, and odds ratios (ORs) were compared between radiological diagnosis and CAD across the range lesion severity levels.
Statistical Tests: Diagnostic performance of the two approaches was compared using *z*-tests on TP rates, FP rates, and the logarithm of ORs across severity levels. A *P*-value <0.05 was considered statistically significant.
Results: ORs of identifying pathology were significantly higher for CAD vis-à-vis visual inspection for all lesions' severity levels. For a 6% change in T_2 value (lowest severity), radiologists' TP and FP rates were not significantly different ($P = 0.12$), while the corresponding CAD results remained statistically significant.
Data Conclusion: CAD is capable of detecting the presence or absence of more subtle lesions with greater precision than the representative group of 25 radiologists chosen in this study.
Level of Evidence: 1
Technical Efficacy: Stage 3

J. MAGN. RESON. IMAGING 2022.

View this article online at wileyonlinelibrary.com. DOI: 10.1002/jmri.28559

Received Aug 30, 2022, Accepted for publication Nov 28, 2022.

*Address reprint requests to: N.B.E., Multidisciplinary Building (203), Suite 406, Haim Levanon 30, Ramat Aviv, Tel Aviv 69978, Israel. E-mail: noambe@tauex.tau.ac.il
The first two authors contributed equally to this work.

From the [1]Department of Biomedical Engineering, Tel-Aviv University, Tel-Aviv, Israel; [2]Sackler School of Medicine, Tel-Aviv University, Tel-Aviv, Israel; [3]Department of Diagnostic Imaging, Sheba Medical Center, Ramat-Gan, Israel; [4]Brain Imaging Research Center (BIRC), Ben-Gurion University, Beer-Sheva, Israel; [5]Zlotowski Center for Neuroscience, Ben-Gurion University of the Negev, Beer-Sheva, Israel; [6]Center for Advanced Imaging Innovation and Research (CAI2R), New York University, New York, New York, USA; and [7]Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel

Additional supporting information may be found in the online version of this article

Magnetic resonance imaging (MRI) is the often preferred modality for noninvasive imaging of soft-tissue pathologies. Traditionally, MRI diagnosis is performed via visual interpretation of contrast-weighted images with typical voxel sizes in the range of ~1–10 mm^3.[1] Most pathologies, however, emerge at the microscopic level and manifest radiologically only after reaching a certain level of severity, for example, multiple sclerosis (MS),[2] Parkinson's disease[3] Alzheimer's disease,[4] or liver metastases.[5] MRI-based diagnosis is thus limited to abnormalities that spread over large enough volumes and above certain levels of severity in order to be visually perceptible. As a result, considerable effort has been invested throughout the last few decades in developing more sensitive tools for earlier detection of tissue abnormalities and more precise assessment of treatment response.[6]

Typical MR images are weighted by one or several MR properties of the tissue, for example, relaxation times,[7] while also being affected by external factors such as the receive or transmit coils sensitivity profiles, and inhomogeneity of the main magnetic field.[8] In the last decade, quantitative MRI (qMRI) techniques have been developed, in which physical parameters responsible for image contrast are determined on a pixel-by-pixel basis to produce parametric maps.[9] These maps provide information pertaining to the tissue's microstructural architecture and chemical or biological composition, which, in turn, correspond to pathological processes.[10] An advantage of qMRI is its improved sensitivity to tissue changes as exemplified by its ability to detect subtle pathology in tissues that look normal under visual inspection.[11] A second advantage of qMRI is its potential to produce values that are invariant across scanners and scan setting, thereby facilitating longitudinal studies and data sharing between medical centers.[11,12] These advantages of qMRI have been demonstrated in various clinical applications, including cardiac, neurologic, and musculoskeletal applications.[11,13,14] Recently, several initiatives have been established, aiming to facilitate and advance the use of qMRI in the clinic (eg, by the Radiological Society of North America, and by the European Society of Radiology[15,16]). Some of these include the use of computer-aided diagnosis (CAD) of parametric qMRI maps as a supplementary approach for visual interpretation of MR images. These include machine learning tools,[17] voxel-based analysis,[18] region-of-interest analysis,[11] and medical decision support systems.[19] CAD is also applied to contrast-weighted image data, albeit with lower robustness to data normalization, scaling, type of scanner, and scan parameters.[20]

One of the key elements for assessing the utility of CAD tools is to test whether they can improve the sensitivity of radiologic readings. This sensitivity can be tested with respect to various features of the tissue pathology such as size, location, or severity of lesion. Several studies exist for evaluating the sensitivity of radiologic reading,[21,22] or comparing human visual analysis vis-à-vis CAD algorithms where ground truth is obtained using other methods (eg, retrospective diagnosis).[17,23–25] One possible disease model for such comparison is MS, which is characterized by inflammatory and demyelinating white matter (WM) lesions, manifesting as hyperintensities on T_2-weighted MR images (see Fig. 1).[2]

Diagnosis of MS is based on the McDonald criteria, which, amongst other parameters, relies on visual estimation of lesion load.[26] Previous studies have shown that CAD based on quantitative mapping of T_2 relaxation times can provide additional useful biomarkers for distinguishing MS patients from healthy controls.[11,18,27]

The aim of this study was thus to compare conventional (visual) MRI assessment of tissue pathology to CAD using a

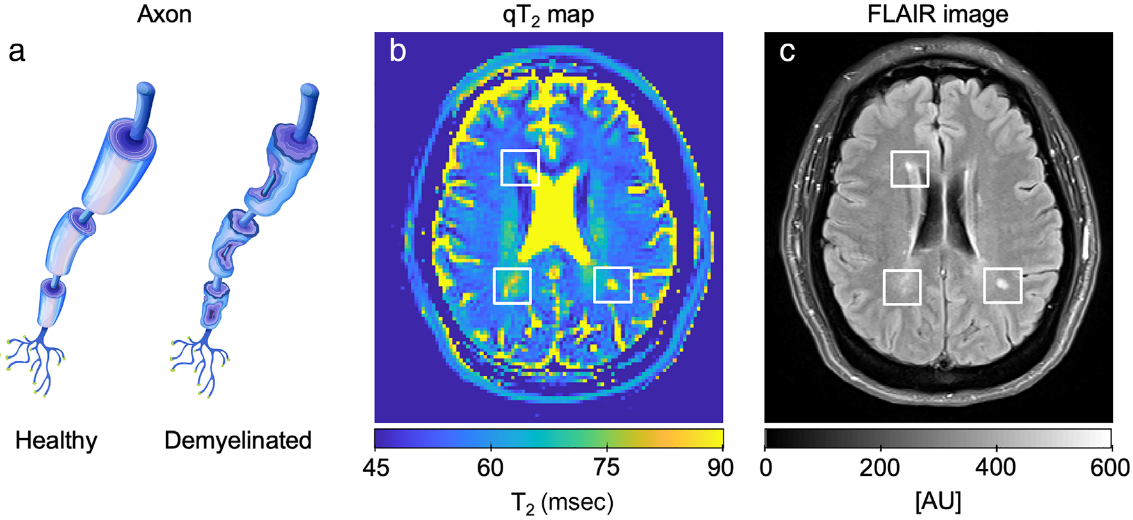


Figure 1: Demyelination processes, their effect on the T_2 relaxation time, and appearance in T_2-weighted fluid attenuated inversion recovery (FLAIR) images. (a) Inflammation leads to demyelination of axons. (b) Inflamed areas are characterized with elevated T_2 values. (c) Regions of elevated T_2 appear as hyperintensities in FLAIR images. Representative lesions are emphasized by white bounding boxes.

deep learning neural network, in the setting of MS diagnosis using simulated T_2-weighted images.

Materials and Methods

Data Collection

Twenty five neuroradiologists (15 males, 15 in training), 29 to 65 years old (mean 39 ± 9), with 1–35 years of experience (mean 9 ± 9.8) were recruited from two large hospitals to participate in the experiment. The experiment was approved by the local institutional review board (IRB no. 0002297-3). Informed consent was obtained from all radiologists who participated in this study. All participants had normal or corrected-to-normal vision.

MRI data from 44 healthy human volunteers was collected after obtaining informed consent and under the approval of the local institutional review board (IRB no. 0001368-1 and S15-00023). One patient was excluded from the study due to a single enhancing lesion found in their hemispheric WM. Scans were performed on whole-body 3 T scanners (Prisma and Skyra, Siemens Healthineers, Erlangen, Germany). Scans used magnetization prepared rapid gradient echo (MPRAGE; Fig. 2a), multi-echo spin-echo (MESE; Fig. 2b) and fluid attenuated inversion recovery (FLAIR) sequences. Scan parameters are given in Table S1 in the Supplemental Material. FLAIR scans from an additional 30 MS patients were imported from the Multiple Sclerosis dataset of the University Hospital of Ljubljana (MSLUB) for pre-training the neural network used for CAD.[28]

Image Analysis

MESE data were used to generate quantitative T_2 (qT_2; Fig. 2d) and proton density (PD; Fig. 2e) maps using the echo modulation curve (EMC) algorithm.[8] WM masks were automatically generated from MPRAGE scans and registered to qT_2 and PD maps using Freesurfer software tools (surfer.nmr.mgh.harvard.edu)[29,30] (Fig. 2c).

Synthetic lesions were embedded into qT_2 maps of healthy brains. Lesions' location and shape were determined using classic image processing tools. First, a focal point was randomly selected within the WM mask. Voxels within a radius of 5 mm around the focal point were then chosen randomly, and their convex hull determined the lesion's area (Fig. 2f). Lesions' with size smaller than 0.5 cm^2 were dilated until their area exceeded 0.5 cm^2, producing lesions of relatively fixed size. MS pathology was simulated by elevating qT_2 values within the lesion's area to one of eight predetermined severity levels: 6%, 9%, 12%, 15%, 18%, 21%, 25%, and 30%. Elevation of values was applied in a spatially centric manner where the increase was maximal at the lesion center, and zero at the edges. Examples of simulated lesions are shown in Fig. 3.

Modified qT_2 maps (Fig. 2g) were used to generate synthetic FLAIR images used in the psychophysical experiment (Fig. 2h). Conversion was performed using an analytic model for the acquisition of FLAIR signal on an MRI scanner (see Eq. S1, Fig. S1 in the Supplemental Material). Model parameters were optimized to visually resemble corresponding acquired FLAIR scans. The synthetic FLAIR images were examined by a neuroradiologist

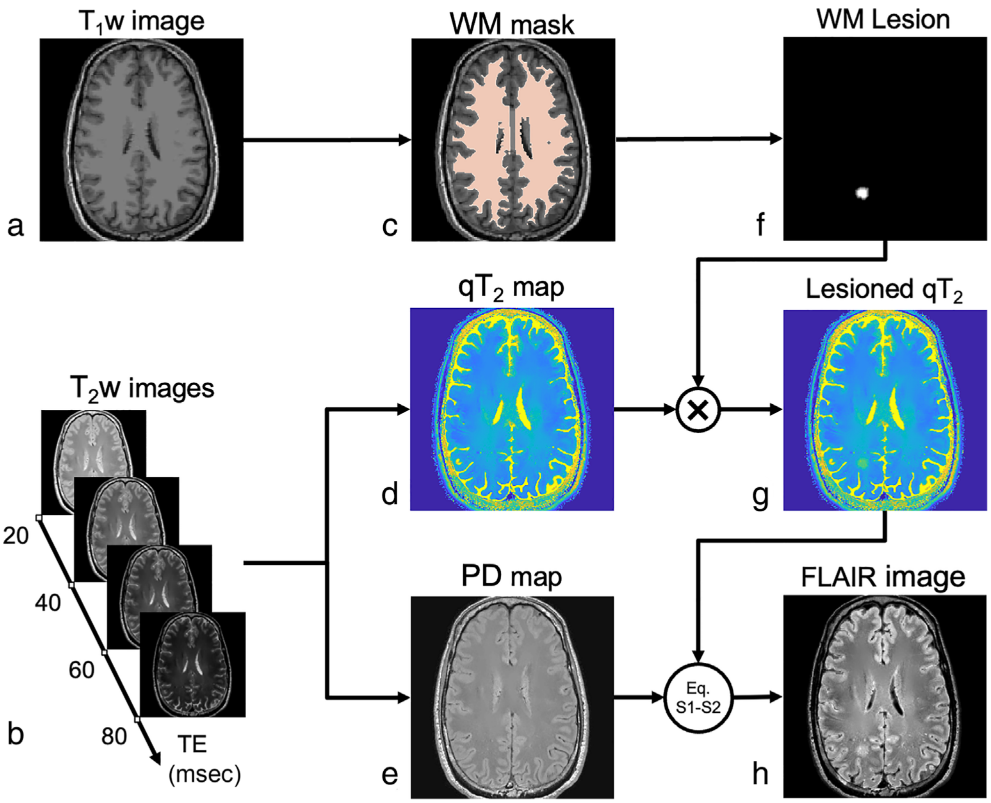


Figure 2: Synthetic magnetic resonance imaging pipeline: (a) T_1-weighted image from a healthy subject. (b) T_2-weighted scans for increasing TEs from a healthy subject. (c) White matter (WM) segmentation generated automatically using Freesurfer software. (d) T_2 map generated using the echo modulation curve (EMC) algorithm.[8] (e) Proton density (PD) map generated using the EMC algorithm. (f) Randomization of a convex ROI in the WM, whose values dictate pathological T_2 changes. (g) Lesioned T_2 map is generated by voxel-wise multiplication of the ROI and the T_2 map. (h) T_2-FLAIR image is synthesized using an analytical signal model.

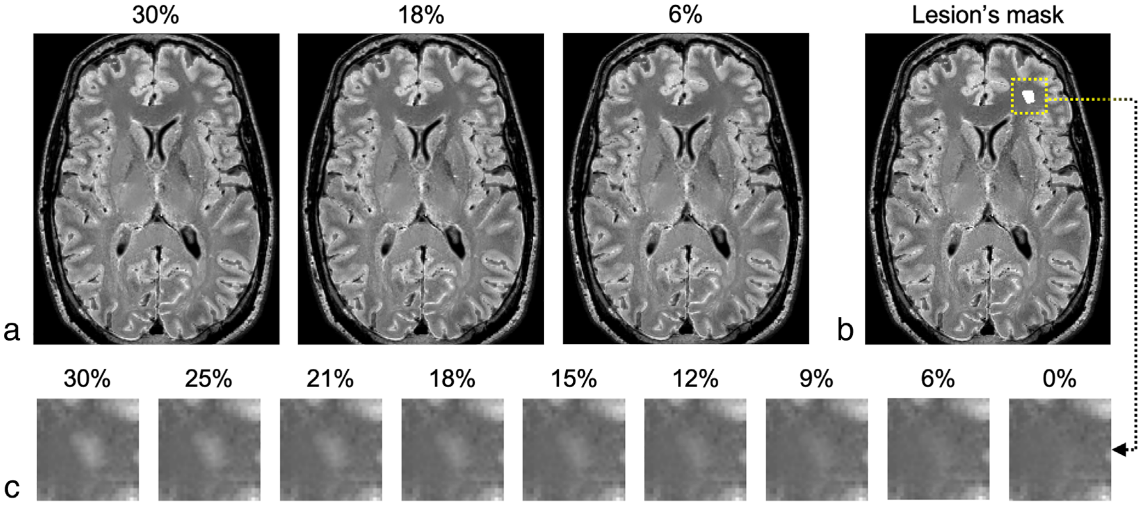


Figure 3: Synthetic lesion embedded on a two-dimensional FLAIR image. A lesion is synthesized by changing the underlying values of the tissue's T_2 relaxation times. (a) Examples of synthetic lesion at three levels of severity, reflecting T_2 changes of 30%, 18%, and 6%. (b) Synthetic lesion in a randomly chosen WM region highlighted in white overlay and a dashed yellow inset. (c) Zoomed view of the lesion in (a) and (b) for nine severity levels, where 0% change indicates a healthy tissue.

(S.S., with 10 years of experience), confirming the appearance of the synthetic FLAIR contrast and the appearance of the simulated MS lesions.

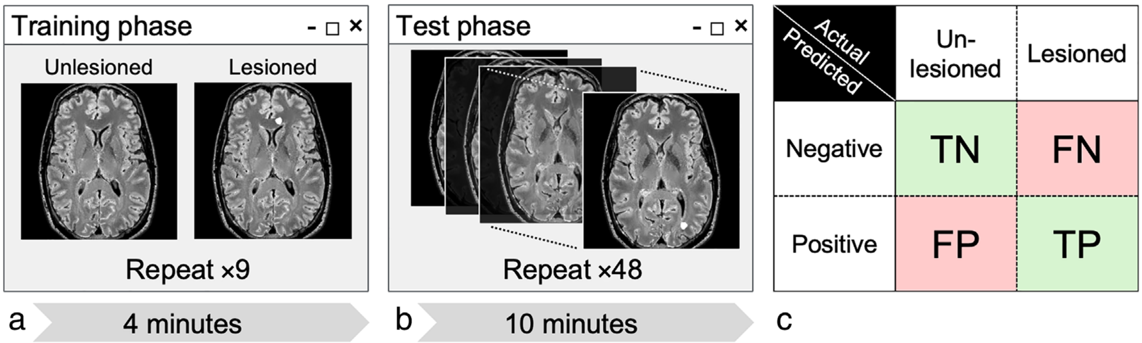
Psychophysical Experiment for Lesion Detection

A two alternative forced choice psychophysical experiment was designed to prospectively assess the efficiency of conventional visual detection of tissue pathology.[31] Stimuli for the experiment consisted of 48 two-dimensional synthetic FLAIR images, derived from scans of three healthy volunteers (dataset 1, see Table S1 in the Supplemental Material), having a constant spatial resolution (256 × 256), and taken from the supratentorial brain region. The image series contained 32 images with a single, oval, hyperintense lesion of similar size, and 16 images that were unedited and lesion free.

A diagram of the psychophysical experiment protocol is illustrated in Fig. 4. Prior to the experiment participants were informed of the relative number of lesioned images. The experiment began with a practice phase consisting of nine images, out of which six were lesioned, while feedback on the detection accuracy was provided for each image. The nine practice scans were used exclusively for training and not for assessment of performance. The actual psychophysical experiment was performed after the practice phase: participants were shown the series of 48 synthetic FLAIR images and asked to point out lesions. Images were present on the screen for 10 seconds each, while blank images were shown for 400 msec between each FLAIR image to reduce afterimage effect.[32] The test phase was split into two parts, each containing 24 images, and separated by an elective break of 1–5 minutes. Participants were allowed to skip images (i.e., shorten the 10 seconds period), and their selections and response time were recorded.

CAD System

To assess CAD-based detection of lesions, the same series of images used for the psychophysical experiment were used as a test set for a binary classification neural network. The network architecture was inspired from Y-Net[33] with an EfficientNet backbone.[34] The network included attention layers (Fig. S3). The attention weight mask was regularized by an innovative scheme (Eq. S2). Pre training of the network was done using MS patients' FLAIR images from a



Actual / Predicted	Un-lesioned	Lesioned
Negative	TN	FN
Positive	FP	TP

Figure 4: Psychophysical trial scheme: (a) Training phase: nine pairs of images are presented. Two out of three images on the right are lesioned, while the left-hand images show the same slice with no lesion. Lesions are highlighted when found, as shown. (b) Test phase: one image is presented at each step. Two out of three images contained lesions at various severity levels. (c) Raw data illustration in a confusion matrix. Correct classifications (TP—true positive/hit; TN—true negative/correct rejection) are highlighted in green while wrong classifications (FP—false positive/false alarm; FN—false negative/miss) are highlighted in light red.

published dataset.[28] Training and validation were performed using the synthetically generated FLAIR images ($N_{Total} = 9600$), containing images of healthy anatomy ($N_{Healthy} = 3200$) and of synthetic lesions ($N_{Lesions} = 6400$). Training and validation images were derived from scans of 41 healthy volunteers (datasets 2 and 3, see Table S1 in the Supplemental Material). Training and validation sets were separated (35 volunteers' scans for training, six for validation). The entirety of data used for training and validation was not included in the psychophysical test. Code for training and evaluating the model is available at https://github.com/OmerShmuelii/models.

Statistical Analysis

SPSS version 24.0 (IBM) and MATLAB R2018a (MathWorks Inc., Natick, MA, USA) were used to evaluate the performance of CAD-based detection versus conventional visual detection of brain lesions. Logistic regression was used to determine the change in accuracy as a function of radiologists' years of experience and lesion severity (% of change in T_2 values), where diagnostic accuracy is defined as $100 \times (TP + TN)/(TP + TN + FP + FN)$. Cohen's kappa coefficient was used to evaluate the agreement between the two detection methods (radiologists and CAD), where kappa values of 0–0.6 were considered weak, 0.6–0.8 were considered moderate, and 0.8–1 were considered as strong agreement.[35] Cohen's kappa scores were calculated for two classifications of the data: once using binary decision (i.e., lesioned/unlesioned), and second using a four-category classification (i.e., TP, FP, TN, FN). Each of these was calculated separately for each severity level and also globally producing an overall kappa score for all severity levels. Odds ratios (ORs) were calculated for CAD and for radiologist detection and compared between the two approaches using z-test for log(OR). A P-value <0.05 was considered statistically significant.

Results

True Positive and False Positive Rates

Participants took an average of 5.6 ± 3.4 seconds to analyze each image, while the overall duration of the psychophysical experiment was 7:42 ± 1:26 minutes. Fig. 5 presents the efficiency of radiologic and of computer-aided detection of lesions, including true positive (TP) and false positive (FP) rates per severity level. Variability in subjective assessments among different radiologists and for CAD are indicated by the error bars. TP rates (solid lines) for both radiologists and CAD increase with lesion severity. The TP rate for CAD was significantly higher than that of radiologists at middle-low severity levels of 9–15% elevation in T_2, but not significantly different at 6% and at levels of 18–30% ($P = 0.07$, 0.80, 0.40, 0.38, and 0.29 for T_2 elevation of 6, 18, 21, 25, and 30% respectively). The FP rate (dashed line) for CAD was significantly lower than that of radiologists (35 of 400 vs. 93 of 400 respectively, i.e., 8.75% vs. 23.3%). At very low severity level (6% elevation in T_2), the radiologists TP rate was not significantly different from their FP rates ($P = 0.12$), indicating that lesions were identified at chance level.

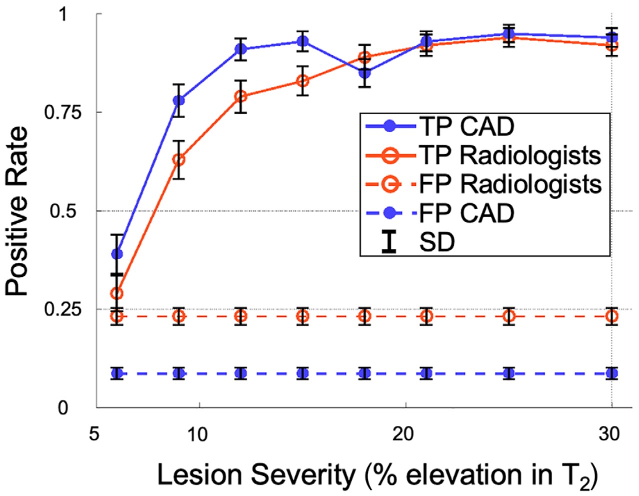


Figure 5: TP and FP rates for radiologic and CAD lesion detection as a function of the lesion severity. TP rate for CAD was significantly higher than that of radiologists at middle-low severity levels of 9%–15% elevation in T_2, and comparable at higher and lower levels. FP rate (dashed line) for CAD is significantly lower than that of radiologists.

Agreement Between Radiologists and CAD

Full list of Cohen's kappa (κ) scores, along with detection performance for radiologists and CAD, is shown in Tables S2 and S3 in the Supplemental Material. Kappa scores for agreement between visual and computer-aided detection indicated weak agreement between the two approaches across all lesion severities. Specific kappa scores were calculated for several classifications of the data: considering only positive / negative binary decision produced a $\kappa = 0.41$; considering a binary decision but separately for each severity level produced kappa scores of $\kappa < 0.48$; considering a four-way diagnosis (TP, TN, FP, FN) across all severity levels, resulted in a score of $\kappa = 0.52$; and lastly, considering a four-way diagnosis but separately for each severity level, produced $\kappa < 0.55$.

ORs Comparison Per Severity Level

Overall ORs for radiologists and for CAD were 11 (95% confidence interval [CI]: 9.5–14) and 53 (95% CI: 42–66) respectively. The CAD overall OR was significantly higher than the radiologists' OR. Figure 6 presents the OR values for radiologists and for CAD across the lesions' severity levels. Variability in subjective assessments among different radiologists and for CAD is indicated by the confidence bars. Analyzing each severity level separately, ORs for CAD were significantly higher than ORs for radiologists for all severity levels (6%, 9%, 12%, 15%, 18%, 21%, 25%, and 30% elevation in T_2). Notably, the radiologists OR for the first severity level (6% elevation in T_2) was not statistically significantly higher than 1. This is consistent with the similarity between the FP and TP rates for severity level of 6% (93 of 400 and 29 of 100 respectively, i.e., 23.3% FP and 29% TP; see Fig. 5), indicating that this level of severity is below the

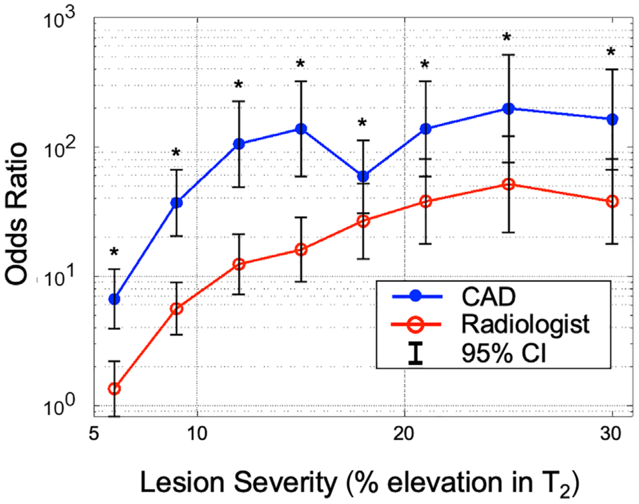


Figure 6: Odds ratios (ORs) for radiologists and computer-aided diagnosis (CAD) as a function of the lesion severity. Error bars indicate 95% confidence intervals (CIs). ORs for both techniques increase with lesion severity. ORs for CAD are significantly higher than ORs for radiologists in the four lowest severity levels (≤15% elevation in T_2 relaxation times) and are comparable for higher lesion severity. *Statistically significant difference: $P < 0.05$ with z-test for log(OR).

threshold of visual detection in the experimental settings of this study.

Trends in Radiologists' Accuracy Per Years of Experience and Per Severity Level

Based on the regression model the radiologists' error rate decreased by 1.9% for each year of experience and decreased by 2.5% per 1% elevation in T_2. Both findings were statistically significant.

Discussion

This work compared the diagnostic performance of radiologists and of a neural-network-based CAD. To that end, a diagnostic psychophysical test was designed for radiologists, and later given as input to the CAD tool. Results showed that the selected CAD tool outperformed radiologists at low lesion severity levels, while providing comparable diagnostic capability at high severity levels. This suggests that CAD has the potential to serve as a guide to radiologic analysis, particularly for early diagnosis of subtle tissue abnormalities.

The psychophysical experiment performed in this study was designed to match clinical settings as closely as possible. This included simulating realistic lesions on standard FLAIR MRI images and authenticating their morphology and location through visual inspection by a neuroradiologist with 10 years of experience. The experimental procedure was also adjusted to maximize similarities with clinical routine, while maintaining a relatively simple binary detection task stating either the existence or the lack of a lesion in each image. An important difference between this work and past studies where radiologists were presented with detection tasks,[21] is that in the previous studies, radiologists rated lesions on a certainty scale, rather than providing a binary decision. The current study did not take such an approach to prevent complications when comparing radiologic diagnosis and CAD. Derivation of certainty levels which are consistent for both radiologists and deep neural networks (DNNs) is possible but requires different experimental design and is thus left for future studies.

The pathologies used as stimuli for the psychophysical experiment consisted of simulated lesions, allowing precise regulation of their location, size, and severity (see the Image Analysis section of the Materials and Methods). The benefit of using synthetic data when training DNNs to detect real MS lesions in MR images, was investigated by Shmueli et al,[36] showing an improvement from 87.5% to 91.2% in a network's accuracy when employing this type of data augmentation. Simulated pathologies in this study were generated at subtle severity levels of 6%–30% elevation in T_2. This choice of severity levels was based on results from a previous study,[11] in which WM lesions that were obvious to neuroradiologists on T_2-weighted images manifested much higher elevation of ≥35% change in T_2 values compared to homologues regions in healthy controls.

In this study, logistic regression and kappa score calculations served as sanity tests for the psychophysical experiment. Kappa scores suggested that the errors made by the radiologists and by the DNN were largely independent from one another and resulted from detection mismatches and not from internal bias of the data—particularly at low severity levels that correspond to early pathology. Trends from the logistic regression analysis were as expected, indicating that more severe lesions were easier to detect, and that more experienced radiologists performed more accurate diagnosis.

OR analysis resulted with expected trends, where ORs for both CAD and radiologists were monotonic functions of the lesion severity, except for a single CAD OR value at severity level of 18%, which we consider an outlier of the experiment. The lack of significant difference between TP and FP rates for radiologists at the lowest severity levels implies that the average evaluator has similar probability of classifying the image as lesioned or unlesioned, regardless of the underlying ground truth. This indicates that at the lowest severity level, radiologic diagnosis was done at a chance level. Furthermore, the analysis indicated that the CAD-based approach outperformed conventional radiologic detection across all severity levels in terms of OR, and at middle-low severity levels (9%–15% elevation in T_2) in terms of TP rates. This suggests that automation of detection tasks may enable more precise and early diagnosis. Alternatively, CAD may be used as a decision-support or triaging systems as was suggested in recent reports.[19,37] Although CAD is unlikely to outperform radiologists on every case, it is more scalable than visual analysis when facing large amounts of data.[38] This implies that experts' time can be saved by embedding new, automated, tools for detecting abnormalities in medical images. Furthermore, CAD has potential to improve the availability of healthcare in specialty fields and in countries with limited number of expert physicians.[37]

The psychophysical experiment used in this study employed a software package implemented in-house, that automatically generated and embedded synthetic lesions in MR images. This platform could be further utilized in several other applications including evaluation of experts from different backgrounds, and deployment as part of radiologists' training programs. Another promising application is to harness this platform to augment data when training machine- and deep-learning based CAD tools.[39] A provisional patent application was filed for this technology in the United States in July 2021 (application number 63/218,414).

Study Limitations

The stimuli used in the experiment had two degrees of freedom: location and severity, while lesion size remained relatively constant. Data analysis, however, was based on a binary decision stating either the existence or the lack of a lesion in each image. This means that an image could have been potentially tagged as having a lesion, yet, in the wrong location. Lesions' location is also relevant when analyzing a binary decision, as it can influence detectability. Incorporating lesions' location or severity in the psychophysical experiment, however, would require a considerably larger number of trials, and more extensive use of experts' time. Moreover, radiologic diagnosis, and particularly differential diagnosis, is more complex than a simple binary detection of a single lesion in a two-dimensional image, as it typically requires addressing the existence of several lesions, in multiple slices and contrasts, while also incorporating the patient's medical history.[26] Another limitation lies in the use of synthetic MS lesions as a model. While this allowed rigorous and accurate investigation of the level of detectability in isolation of other cofactors, it was at the cost of a simplified disease model. Improvements to this model may be achieved by generalizing the pathological tissue changes, for example, by employing multiple lesions, multi-component T_2 distributions,[40] changes in T_1 values,[18] nonconvex morphologies, or completely different pathologies like hepatic lesions, spinal cord injuries, or occult pathology in normal appearing tissues.[11] We thus limit the interpretation of our results for synthetic MS lesions. Further validation should be done on actual lesions and other general pathologies.

Conclusion

This study of simulated MS lesions demonstrated that CAD outperformed radiologists at low-severity lesions and achieved similar performance for moderate to high severities.

Acknowledgments

This work was supported by the Israel Science Foundation (grant no. 2009/17). We would like to thank Itzik Cohen for his consultation on statistics.

Conflict of Interest

Noam Ben-Eliezer and Chen Solomon submitted a provisional patent application for synthesizing pathology in MR images using quantitative MRI signal model (US provisional patent application number 63/218,414). Other authors have nothing to disclose.

Data Availability Statement

Data generated or analyzed during the study are available from the corresponding author by request under restrictions of patient's confidentiality.

References

1. Lüsebrink F, Sciarra A, Mattern H, Yakupov R, Speck O. T1-weighted in vivo human whole brain MRI dataset with an ultrahigh isotropic resolution of 250 μm. Sci Data 2017;4:170032.
2. Reich DS, Lucchinetti CF, Calabresi PA. Multiple sclerosis. N Engl J Med 2018;378:169-180.
3. Poewe W, Seppi K, Tanner CM, et al. Parkinson disease. Nat Rev Dis Primers 2017;3:17013.
4. Johnson KA, Fox NC, Sperling RA, Klunk WE. Brain imaging in Alzheimer disease. Cold Spring Harb Perspect Med 2012;2:a006213.
5. Hardie AD, Naik M, Hecht EM, et al. Diagnosis of liver metastases: Value of diffusion-weighted MRI compared with gadolinium-enhanced MRI. Eur Radiol 2010;20:1431-1441.
6. Harisinghani MG, O'Shea A, Weissleder R. Advances in clinical MRI technology. Sci Transl Med 2019;11:eaba2591.
7. Bloch F. Nuclear induction. Phys Rev 1946;70:460-474.
8. Ben-Eliezer N, Sodickson DK, Block KT. Rapid and accurate T 2 mapping from multi-spin-echo data using Bloch-simulation-based reconstruction. Magn Reson Med 2015;73:809-817.
9. Ben-Eliezer N. Advances in signal processing for relaxometry. In: Nicole S, Vikas G, editors. *Quantitative magnetic resonance imaging*. London, United Kingdom: Elsevier; Vol 1; 2020. p 123-147.
10. Weiskopf N, Edwards LJ, Helms G, Mohammadi S, Kirilina E. Quantitative magnetic resonance imaging of brain anatomy and in vivo histology. Nat Rev Phys 2021;3:570-588.
11. Shepherd TM, Kirov II, Charlson E, et al. New rapid, accurate T2 quantification detects pathology in normal-appearing brain regions of relapsing-remitting MS patients. Neuroimage Clin 2017;14:363-370.
12. Gracien R-M, Maiworm M, Brüche N, et al. How stable is quantitative MRI?—Assessment of intra- and inter-scanner-model reproducibility using identical acquisition sequences and data analysis programs. Neuroimage 2020;207:116364.
13. Messroghli DR, Moon JC, Ferreira VM, et al. Clinical recommendations for cardiovascular magnetic resonance mapping of T1, T2, T2* and extracellular volume: A consensus statement by the Society for Cardiovascular Magnetic Resonance (SCMR) endorsed by the European Association for Cardiovascular Imaging (EACVI). J Cardiovasc Magn Reson 2017;19:75.
14. Ben-Eliezer N, Raya JG, Babb JS, Youm T, Sodickson DK, Lattanzi R. A new method for cartilage evaluation in femoroacetabular impingement using quantitative T2 magnetic resonance imaging: Preliminary

validation against arthroscopic findings. Cartilage 2021;13(1_suppl): 1315 S-1323 S.

15. Kessler LG, Barnhart HX, Buckler AJ, et al. The emerging science of quantitative imaging biomarkers terminology and definitions for scientific studies and regulatory submissions. Stat Methods Med Res 2015; 24:9-26.

16. deSouza NM, Achten E, Alberich-Bayarri A, et al. Validated imaging biomarkers as decision-making tools in clinical trials and routine practice: Current status and recommendations from the EIBALL* subcommittee of the European Society of Radiology (ESR). Insights Imaging 2019;10:87.

17. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. Lancet Digit Health 2019;1:e271-e297.

18. Piredda GF, Hilbert T, Granziera C, et al. Quantitative brain relaxation atlases for personalized detection and characterization of brain pathology. Magn Reson Med 2020;83:337-351.

19. Seah JCY, Tang CHM, Buchlak QD, et al. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: A retrospective, multireader multicase study. Lancet Digit Health 2021;3:e496-e506.

20. Schurink NW, van Kranen SR, Roberti S, et al. Sources of variation in multicenter rectal MRI data and their effect on radiomics feature reproducibility. Eur Radiol 2022;32:1506-1516.

21. Woo JH, Henry LP, Krejza J, Melhem ER. Detection of simulated multiple sclerosis lesions on T2-weighted and FLAIR images of the brain: Observer performance. Radiology 2006;241:206-212.

22. Altay EE, Fisher E, Jones SE, Hara-Cleaver C, Lee J-C, Rudick RA. Reliability of classifying multiple sclerosis disease activity using magnetic resonance imaging in a multiple sclerosis clinic. JAMA Neurol 2013;70: 338-344.

23. Freeman K, Geppert J, Stinton C, Todkill D, Johnson S, Clarke A, et al. Use of artificial intelligence for image analysis in breast cancer screening programmes: Systematic review of test accuracy. BMJ 2021;374: n1872. https://doi.org/10.1136/bmj.n1872

24. Bilello M, Arkuszewski M, Nasrallah I, Wu X, Erus G, Krejza J. Multiple sclerosis lesions in the brain: Computer-assisted assessment of lesion load dynamics on 3D FLAIR MR images. Neuroradiol J 2012;25:17-21.

25. Bilello M, Arkuszewski M, Nucifora P, et al. Multiple sclerosis: Identification of temporal changes in brain lesions with computer-assisted detection software. Neuroradiol J 2013;26:143-150.

26. Thompson AJ, Banwell BL, Barkhof F, et al. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. Lancet Neurol 2018;17: 162-173.

27. Hagiwara A, Hori M, Yokoyama K, et al. Utility of a multiparametric quantitative MRI model that assesses myelin and edema for evaluating plaques, Periplaque white matter, and normal-appearing white matter in patients with multiple sclerosis: A feasibility study. Am J Neuroradiol 2017;38:237-242.

28. Lesjak Ž, Galimzianova A, Koren A, et al. A novel public MR image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater consensus. Neuroinformatics 2018;16:51-63.

29. Fischl B, Salat DH, Busa E, et al. Whole brain segmentation. Neuron 2002;33:341-355.

30. Greve DN, Fischl B. Accurate and robust brain image alignment using boundary-based registration. Neuroimage 2009;48:63-72.

31. Wickens TD. Forced-choice procedures. *Elementary signal detection theory*. Oxford: Oxford University Press; 2001. p 93-112.

32. Daw NW. Why after-images are not seen in normal circumstances. Nature 1962;196:1143-1145.

33. Mehta S, Mercan E, Bartlett J, Weaver D, Elmore J, Shapiro L. Y-Net: Joint segmentation and classification for diagnosis of breast biopsy images. *International conference on medical image computing and computer-assisted intervention*. arXiv; 2018. p 893-901. https://arxiv.org/abs/1806.01313

34. Tan M, Le Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In: Chaudhuri K, Salakhutdinov R, editors. *Proceedings of the 36th international conference on machine learning*. arXiv; Vol 97; 2019. p 6105-6114. https://arxiv.org/abs/1905.11946

35. McHugh ML. Interrater reliability: The kappa statistic. Biochem Med 2012;22:276-282.

36. Shmueli O, Solomon C, Ben-Eliezer N, Greenspan H. Deep learning based multiple sclerosis lesion detection utilizing synthetic data generation and soft attention mechanism. In: Iftekharuddin KM, Drukker K, Mazurowski MA, Lu H, Muramatsu C, Samala RK, editors. *Medical imaging 2022: Computer-aided diagnosis*. San Diego: Computer-Aided Diagnosis; 2022. p 120330R.

37. Dahan A, Wang W, Gaillard F. Computer-aided detection can bridge the skill gap in multiple sclerosis monitoring. J Am Coll Radiol 2018;15: 93-96.

38. Aljabri M, AlAmir M, AlGhamdi M, Abdel-Mottaleb M, Collado-Mesa F. Towards a better understanding of annotation tools for medical imaging: A survey. Multimed Tools Appl 2022;81:25877-25911.

39. Cha KH, Petrick N, Pezeshk A, et al. Evaluation of data augmentation via synthetic images for improved breast mass detection on mammograms using deep learning. J Med Imaging 2019;7:1.

40. MacKay A, Laule C, Vavasour I, Bjarnason T, Kolind S, Mädler B. Insights into brain microstructure from the T2 distribution. Magn Reson Imaging 2006;24:515-525.