AJNR

Deep Learning–Generated Synthetic MR Imaging STIR Spine Images Are Superior in Image Quality and Diagnostically Equivalent to Conventional STIR: A Multicenter, Multireader Trial

L.N. Tanenbaum, S.C. Bash, G. Zaharchuk, A. Shankaranarayanan, R. Chamberlain, M. Wintermark, C. Beaulieu, M. Novick and L. Wang

Deep Learning–Generated Synthetic MR Imaging STIR Spine Images Are Superior in Image Quality and Diagnostically Equivalent to Conventional STIR: A Multicenter, Multireader Trial

L.N. Tanenbaum, S.C. Bash, G. Zaharchuk, A. Shankaranarayanan, R. Chamberlain, M. Wintermark, C. Beaulieu, M. Novick, and L. Wang

ABSTRACT

BACKGROUND AND PURPOSE: Deep learning image reconstruction allows faster MR imaging acquisitions while matching or exceeding the standard of care and can create synthetic images from existing data sets. This multicenter, multireader spine study evaluated the performance of synthetically created STIR compared with acquired STIR.

MATERIALS AND METHODS: From a multicenter, multiscanner data base of 328 clinical cases, a nonreader neuroradiologist randomly selected 110 spine MR imaging studies in 93 patients (sagittal T1, T2, and STIR) and classified them into 5 categories of disease and healthy. A DICOM-based deep learning application generated a synthetically created STIR series from the sagittal T1 and T2 images. Five radiologists (3 neuroradiologists, 1 musculoskeletal radiologist, and 1 general radiologist) rated the STIR quality and classified disease pathology (study 1, $n = 80$). They then assessed the presence or absence of findings typically evaluated with STIR in patients with trauma (study 2, $n = 30$). The readers evaluated studies with either acquired STIR or synthetically created STIR in a blinded and randomized fashion with a 1-month washout period. The interchangeability of acquired STIR and synthetically created STIR was assessed using a noninferiority threshold of 10%.

RESULTS: For classification, there was a decrease in interreader agreement expected by randomly introducing synthetically created STIR of 3.23%. For trauma, there was an overall increase in interreader agreement by +1.9%. The lower bound of confidence for both exceeded the noninferiority threshold, indicating interchangeability of synthetically created STIR with acquired STIR. Both the Wilcoxon signed-rank and t tests showed higher image-quality scores for synthetically created STIR over acquired STIR ($P < .0001$).

CONCLUSIONS: Synthetically created STIR spine MR images were diagnostically interchangeable with acquired STIR, while providing significantly higher image quality, suggesting routine clinical practice potential.

ABBREVIATIONS: Acq-STIR = acquired STIR; CNN = convolutional neural network; DL = deep learning; IQ = image quality; RMSE = root mean square error; RMSPE = root mean square percentage error; Syn-STIR = synthetically created STIR

A typical clinical protocol for spine MR imaging uses T1WI, T2WI, and STIR scans to depict anatomy and provide adequate sensitivity to a variety of pathologic conditions.

STIR offers a combination of T1 and T2 contrast-weighting and nulled fat signal to highlight pathologic changes in tissues. The fat suppression offered by STIR is more uniform and resistant to magnetic field inhomogeneities than other fat-saturation methods such as spectral "fat-sat," especially near metallic foreign bodies, tissue interfaces with high susceptibility differences (like the skull base/sinuses), and across large body parts like the spine. On lower-field permanent magnets with lower homogeneity, STIR may be the only fat-suppression method available. STIR images have inherently lower SNR than T1WI and T2WI. Despite approaches that use larger voxel sizes to mitigate this challenge, scan times are still long, more susceptible to motion, and harder for patients to tolerate.

Deep learning (DL)-based reconstruction techniques mitigate this challenge by enabling faster acquisitions while matching or even exceeding standard-of-care image quality (IQ).[1-4] Recent work has led to DL methods that can generate entirely synthetic image contrasts, potentially shortening overall study times by removing the need to acquire certain series.

Synthesizing new contrast information from available images has been an active area of research in the MR imaging domain. Convolutional neural network (CNN)-based approaches have

demonstrated state-of-the-art performance for MR imaging contrast synthesis.[5-8] While most of the literature has focused on one-to-one synthesis, several studies considered the many-to-one synthesis problem, in which the algorithm takes multiple contrasts as input and generates 1 missing contrast.[5,7,9] Previous work has demonstrated the potential to synthesize STIR images from T1WI and T2WI.[10] This study goes further by using examinations from multiple scanner manufacturers and a wider variety of magnetic field strengths, including a more comprehensive set of pathologies, and by performing voxelwise analysis of the synthetically created STIR (Syn-STIR) images. Furthermore, the technical methods used in this work integrate multiple maps (including an anatomy-aware segmentation map and a pathology saliency map) in the reconstruction network, so the syn-STIR images maintain high consistency with the acquired STIR (Acq-STIR) images. Our methods also avoid the use of generative adversarial networks, which are prone to introducing structures in synthesized images that are not present in the source images.

This multicenter, multireader study evaluated the diagnostic interchangeability and qualitative image quality of a DL-generated Syn-STIR against a clinical standard-of-care Acq-STIR. There are established methods to assess the interchangeability of the 2 image-acquisition methods, which determine whether the images are diagnostically equivalent.[11] Two images are interchangeable or diagnostically equivalent if a given patient would receive the same diagnosis regardless of which of the 2 images was used. Diagnostic equivalence is tested by comparing an interreader agreement using the baseline imaging method with an interreader agreement using the method being tested versus the baseline while accounting for variability across cases and readers. In addition, multiple quantitative methods were also used to compare the Syn-STIR images with the Acq-STIR images.

MATERIALS AND METHODS
Overview
A DL model was applied to synthesize a sagittal Syn-STIR series from the sagittal T1 and sagittal T2 of clinical spine MR imaging studies. The model contains 3 phases, an anatomy-aware map, a pathology-aware map, and a reconstruction map. For the anatomy-aware map, the segmentation map was obtained for each anatomy, making the anatomy-based operation feasible. The pathology-aware map is a saliency map used to guide the network to maintain pathologic consistency. During the training process, the 2 inputs (sagittal T1 and sagittal T2) were fed into the reconstruction network under separate branches and later concatenated to avoid potential blurriness due to misregistration. We implemented the DL model in TensorFlow (https://www.tensorflow.org/), trained on an NVIDIA V100 GPU (https://www.nvidia.com/en-us/data-center/v100/) with an ADAM optimizer (https://machinelearningjourney.com/index.php/2021/01/09/adam-optimizer/),[12] and applied image registration between the 2 inputs to reduce potential misalignment. The network was trained by comparing the output Syn-STIR image with the Acq-STIR image through multiple loss functions (Online Supplemental Data).

Participants and Distribution of Pathologies
With institutional review board approval, a nonreader senior neuroradiologist identified the dominant pathology (as described below) in a multicenter, multiscanner data base of 328 approximately equal numbers of cervical, thoracic, and lumbar spine MR imaging cases referred for a variety of conditions. From this group, 93 unique patients were evaluated in 2 separate studies. First, 80 patients (40 females, 36 males, 4 not available; age range, 16–89 years) were selected randomly from among 5 categories of disease (defined as the most dominant pathology) based on the findings on the complete study (study 1). The categories were cord lesion ($n = 8$), noncord lesion ($n = 15$), degenerative disease ($n = 20$), infection ($n = 10$), trauma ($n = 17$), and healthy ($n = 10$). The readers were given instructions outlining which clinical entities should fall into each category and to help with classification when multiple pathologies were present. More details can be found in the Online Supplemental Data. In addition, a second study evaluating the ability of readers to identify important features in the setting of trauma was performed (study 2). Patients (13 men, 17 women; age range, 18–89 years) for study 2 included 10 with no imaging evidence of trauma (separate from the patients in study 1) and 20 with imaging evidence of trauma (17 of the patients with trauma in study 1 supplemented by 3 additional patients). These cases were evaluated for the following findings: prevertebral fluid collections (class I), bone edema related to fracture (class II), and posterior soft-tissue/ligamentous injury (class III). The case distribution was class I/II ($n = 3$), class I/II/III ($n = 7$), class I/III ($n = 5$), class II ($n = 1$), class II/III ($n = 1$), class III ($n = 4$), and class none ($n = 9$).

Image Acquisition
The images were acquired on a variety of scanners, including 3T Discovery 750 and 750w, 3T Signa Premier, and 1.5T HDxt (GE Healthcare); 3T Magnetom Skyra, 3T MagnetomVerio (Siemens); 1.5T Intera (Philips Healthcare); 1.5T Vantage Titan (Canon); 0.6T (Fonar Upright) and 0.3T AIRIS Elite (Hitachi/Fujifilm). The case distribution by field strength for study 1 was 0.6T ($n = 1$), 1T ($n = 1$), 1.5T ($n = 43$), and 3T ($n = 35$), and for study 2, it was 1T ($n = 2$), 1.5T ($n = 16$), and 3T ($n = 12$). The image acquisitions consisted of sagittal T1, T2, and STIR series using the individual institution's routine clinical protocol. Section thickness ranged from 3 to 5 mm. FOV varied from 18 to 24 cm cervical; 27 to 30 cm lumbar; and 30 to 38 cm thoracic; and the acquisition matrix varied from 192×192 to 800×380.

Image Processing
The Syn-STIR images were created off-line from existing DICOM images using a vendor-neutral, CNN software application (SubtleSYNTH; Subtle Medical). The CNN was trained to generate synthetic sagittal STIR images using the sagittal T1 and T2 images as input. Because the application was DICOM-based, processing did not require proprietary raw k-space input; thus, it was capable of processing images from any MR imaging platform. The training set included hundreds of thousands of MR images from a variety of vendors, scanner models, field strengths, and clinical sites, as well as a variety of disease states/clinical
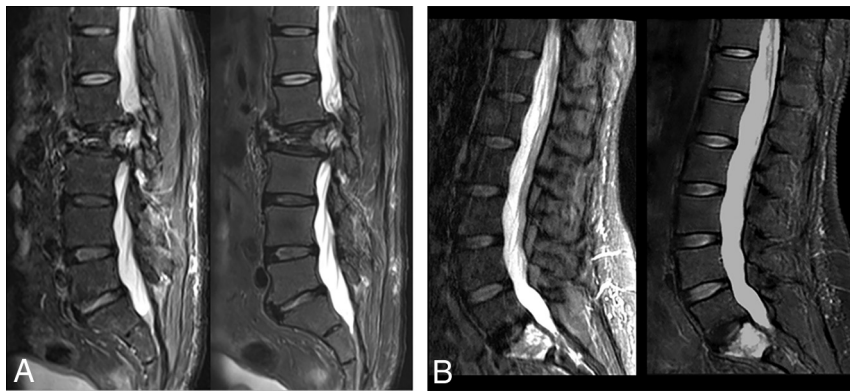
FIG 1. *A*, L2 body fracture. Conventional acquired STIR (*left*) and DL synthesized STIR (*right*). Note the Syn-STIR manifests fewer motion artifacts, improved sharpness, and greater apparent SNR. *B*, Sacral metastasis. Conventional Acq-STIR (*left*) and DL Syn-STIR (*right*). Note the striking reduction in motion artifacts as well as superior overall quality of the Syn-STIR.

indications, thus experiencing a range of tissue contrasts, acquisition parameters, patient anatomies, and image quality.

Image Assessment

Five radiologists (3 neuroradiologists, 1 musculoskeletal radiologist, 1 general radiologist experienced in spine MR imaging) evaluated 160 cases. Each case consisted of 3 sagittal image series: either T1, T2, and Acq-STIR ($n = 80$) or T1, T2, and Syn-STIR ($n = 80$). The image sets were presented in a blinded and randomized fashion on a commercial DICOM viewer, with a 1-month washout period between reading sessions (study 1). To assess diagnostic equivalence, each reader individually classified the pathologies present. Readers also rated the Acq-STIR and Syn-STIR image quality on a 5-point Likert scale (1 = unacceptable, 2 = poor, 3 = adequate, 4 = good, 5 = excellent), which served as a collective summary assessment of individual image-quality metrics, such as perceived SNR, contrast-to-noise ratio, image sharpness, and artifacts. The same readers also evaluated the trauma-specific study ($n = 30$ subjects, $n = 60$ studies) in the same blinded and randomized fashion with the same 1-month washout period (study 2). They were asked to individually classify the findings for the presence/absence of the following: 1) prevertebral fluid collections, 2) fracture-related bone edema, and 3) posterior soft-tissue/ligamentous injury. In addition, 2 neuroradiologists, including one not involved with studies 1 or 2, performed a blinded side-by-side, qualitative evaluation of the study 1 Acq-STIR and Syn-STIR images, rating the extent of disease and diagnostic confidence on a 5-point Likert scale, as well as noting evidence or absence of image aberrations.

RESULTS

Diagnostic-Equivalence Analysis

The 2 imaging methods were assessed for interchangeability or diagnostic equivalence by comparing the interreader agreement within Acq-STIR images with the interreader agreement between Syn-STIR and Acq-STIR images. Interreader agreement for Acq-STIR images was calculated as the percentage of comparisons between 2 different readers for the same case in which the readers' classifications agreed. Interreader agreement for Acq-STIR

versus Syn-STIR images was calculated as the percentage of comparisons between 2 different readers for the same case that the readers' classification when 1 reader was using the Acq-STIR image and the other reader was using the Syn-STIR image agreed with each other. The agreement probability was calculated by mean of a logistic regression model with random effects using the "glmer" function from the "lme4" package in R statistical and computing software (http://www.r-project.org/) following methods described in the literature.[11,13] A noninferiority analysis was performed with a preset hypothesis that the difference in diagnostic classification for interreader agreement for Acq-STIR and Syn-STIR was not >10% lower than the interreader agreement between Syn-STIR and Acq-STIR.

Image-Quality Statistical Analysis

Wilcoxon rank-sum tests were performed to assess the equivalence or superiority of the image quality for each feature. Statistically significant superiority for a feature was determined by $P < .05$. Adjustment for significance tests for multiple comparisons was made using a Bonferroni correction.

Voxel-Intensity Analysis

To evaluate the voxelwise correlation between the Syn-STIR image and the conventional STIR image, we drew 4 ROIs on each target tissue (vertebral bone, disc, CSF, spinal cord, and fat) and calculated the mean of the 4 ROIs per series. Note that areas without any pathologies were selected. For example, in patients with degenerative disease, ROIs were drawn only on the healthy disc. Similar rules applied to other tissues as well. Bland-Altman analysis[14,15] was then applied, followed by the Shapiro-Wilk results on the difference.[16] Additionally, root mean square error (RMSE) and root mean square percentage error (RMSPE) were calculated. A Passing-Bablok regression analysis was performed to evaluate agreement between the 2 images.[17]

Sample image pairs are shown in Fig 1, demonstrating similar fat-saturated T2-weighted image contrast of the Syn-STIR as the Acq-STIR. Lower noise levels are seen in the Syn-STIR images.

Diagnostic Interchangeability

The estimate of interchangeability (diagnostic equivalence) when accounting for readers and cases as random effects was −3.23% (95% CI, –6.61%–0.19%), evaluated over 1000 bootstrapped samples (Fig 2). The decrease in interreader agreement expected when interchanging Acq-STIR images with Syn-STIR images was 3.23%. Based on the results, the estimate of interchangeability was not significantly worse than the noninferiority limit of 10% ($P = .001$). On the basis of the prespecified noninferiority criteria of 10%, we concluded that interchanging the Acq-STIR images with Syn-STIR images would not lead to a significant decrease in

interreader agreement; thus, the Syn-STIR was deemed diagnostically equivalent to the Acq-STIR.

For the trauma subset, the 3 structure-based classifications (prevertebral fluid collections, fracture-related bone edema, and posterior soft-tissue/ligamentous injury) were analyzed separately as different classes (Fig 3). The estimate of interchangeability (diagnostic equivalence) when accounting for readers and cases as random effects for the 3 classes was +0.85% (95% CI, −4.13%–5.48%), +2.3% (95% CI, −2.8%–7.1%), and +2.2% (95% CI, −2.2%–6.4%), respectively; each class evaluated >1000 bootstrapped samples. In other words, the interreader agreement can be expected to improve by 0.85%, 2.3%, and 2.2% when interchanging traditional STIR images with Syn-STIR images. Based on the results, the estimate of interchangeability was not significantly worse than the noninferiority limit of 10% (P = .001). Given all 3 classes, a final analysis was performed in which the results described above from the 3 classes were combined, and "class" was included as a fixed effect in the statistical model. The interchangeability estimate was +1.9% (95% CI, −1.1%–5.0%), indicating that there was an improvement in the interreader agreement found when interchanging Acq-STIR images with Syn-STIR images. We, therefore, conclude that for the trauma study Syn-STIR was interchangeable with Acq-STIR.

Image-Quality Analysis
Acq-STIR images had an average IQ score of 3.21 (SD, 1.08), and Syn-STIR images scored an average of 3.71 (SD, 1.14). A Wilcoxon signed-rank test showed a significantly higher median IQ score for Syn-STIR images than Acq-STIR images (median = 0.4, P < .0001). A t test on the paired difference in IQ scores across artificial intelligence–generated and Acq-STIR images showed a significantly higher average
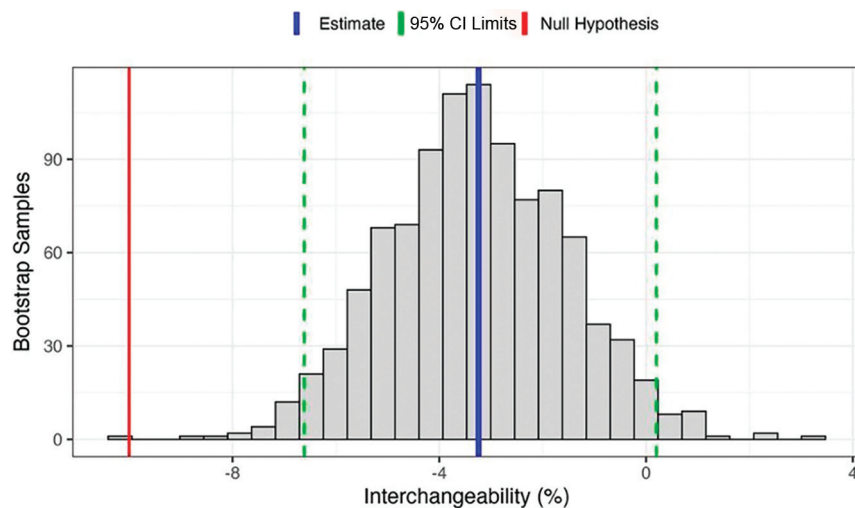

FIG 2. Interchangeability when accounting for readers and cases as random effects under the whole-cohort estimated >1000 bootstrapped samples.
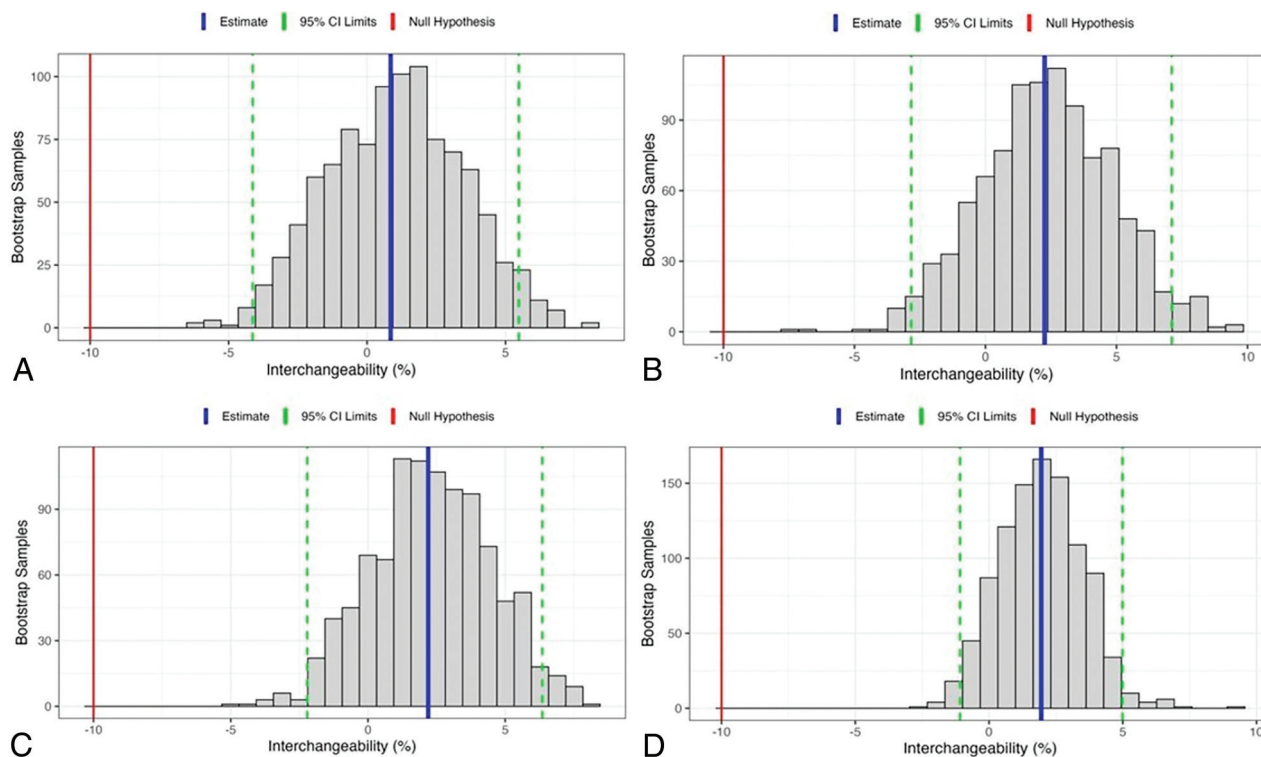

FIG 3. Interchangeability when accounting for readers and cases as random effects under the trauma cohort estimated >1000 bootstrapped samples. The x-axis represents the estimated interchangeabilities in the unit of percentage, and the y-axis represents the count of bootstrapped samples. A, Trauma class I. B, Trauma class II. C, Trauma class III. D, Mixed effect for all 3 classes.
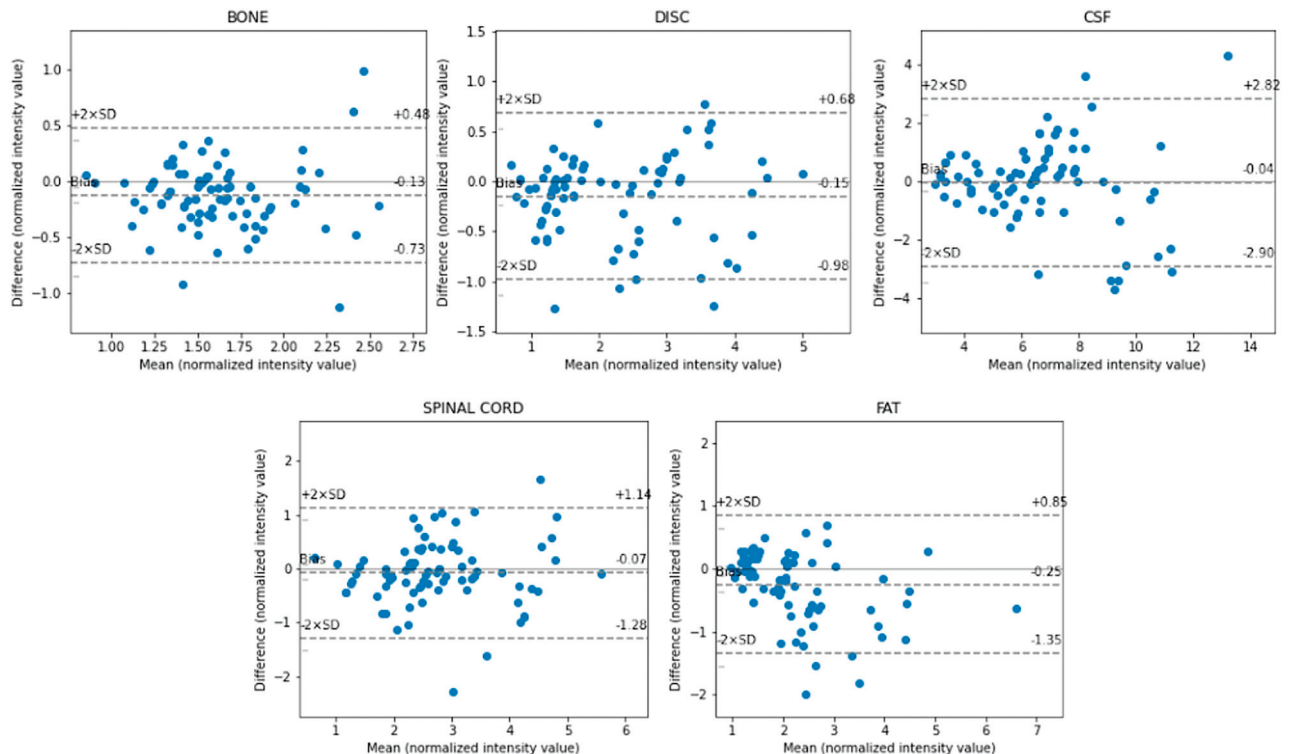
FIG 4. Bland-Altman plots for each tissue from 80 clinical cases. The x-axis represents the mean of the normalized intensity value from Syn-STIR and its nearest neighbor from Acq-STIR, and the y-axis represents the difference between the normalized intensity value from Syn-STIR and its nearest neighbor from the Acq-STIR.

IQ score for Syn-STIR images compared with Acq-STIR images (mean paired difference = 0.50; 95% CI, 0.33–0.67; $P < .0001$).

Side-by-Side Comparison
In the blinded, side-by-side evaluation of the cases in study one, 94.9% of Syn-STIR sets demonstrated equal or a better extent of disease compared with Acq-STIR for the first reader and 97.5% for the second reader; 88.6% of cases provided equal or higher diagnostic confidence with Syn-STIR for the first reader and 87.3% for the second reader. In addition, no unexpected differences were found between the 2 STIR types, indicating that the Syn-STIR method did not create unique artifacts.

Voxel Consistency
The Bland-Altman plots for voxel consistency are shown in Fig 4. For each tissue, the bias (the mean of the difference between the Acq-STIR and Syn-STIR) was close to zero. The smallest average bias was from the CSF, which was −0.04 normalized intensity units, and the largest average bias was from fat, which was around −0.25 normalized intensity units. The Shapiro-Wilk results showed that all P values were $> .05$, implying that the difference between the Acq-STIR and the Syn-STIR is normally distributed.

In addition, the RMSE and RMSPE between the Acq-STIR and Syn-STIR images for each patient were 0.45 and 17.88 normalized intensity units, respectively. For all 80 cases, the median of the RMSE value was 0.45 normalized intensity units, and the median of the RMSPE percentage was 17.9%. After confirming that the 5 tissues passed the Shapiro-Wilk test for normality, the Passing-Bablok regression was applied to estimate the regression

line and intercept (Fig 5). The slopes of the disc, CSF, and spinal cord were 1.06, 1.05, and 1.07, which indicate a high correlation between the 2 results. The slope of bone and fat was 0.85 and 0.78, respectively. The results indicate excellent voxelwise consistency between the Acq-STIR and the Syn-STIR images.

DISCUSSION
STIR is quite powerful in depicting spine pathology and thus is part of almost all routine spine imaging protocols; however, conventional reconstruction scan times are long. Also, because of the fat inversion pulse, the SNR of the images is lower than that of other sequences. A synthetically generated STIR could result in approximately 3–5 minutes of scan time avoided or up to 25% overall time-savings per examination, increasing imaging enterprise efficiency. Because up to 30% of patients report significant anxiety, largely from claustrophobia, during an MR imaging study, scan-time reductions inherently improve the patient's experience.[18] The authors' internal multicenter surveys have shown that even minor reductions in examination length result in a significantly higher level of patient satisfaction.[19]

MR imaging examinations are susceptible to image degradation from motion, particularly during lengthy scans. Motion is a significant challenge in MR imaging, occurring in 29% of inpatient/emergency department examinations and 7% of outpatient studies[20] and can lead to the need to repeat sequences or entire studies. Andre et al[21] found that 19.8% of all MR imaging sequences needed to be repeated due to motion artifacts, correlating with US $592 revenue loss per hour and an annual loss of US $115,000 per scanner.
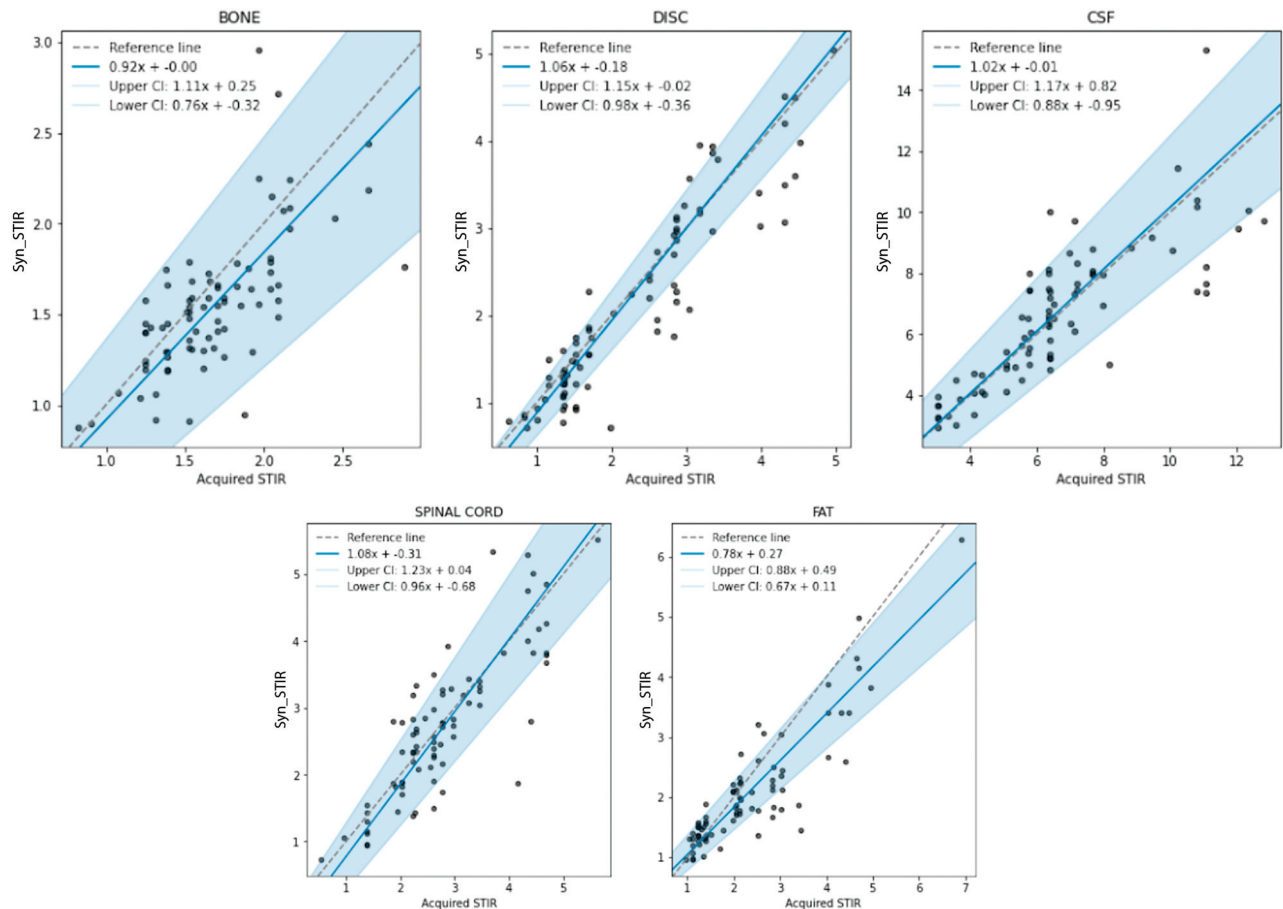
FIG 5. Passing-Bablok regression applied to estimate the regression line and intercept. The slope of the disc, CSF, and spinal cord is 1.06, 1.05, and 1.07, which indicates a high correlation between the 2 results. The slope of bone and fat is 0.85 and 0.78, respectively. The results indicate excellent voxelwise consistency between the Acq-STIR and Syn-STIR images.

The generally inverse relationship between MR image quality and scan duration is well-established.[1,2] Traditionally reconstructed, high-resolution, high-SNR images require acquisition times that can be quite long. DL-based image reconstruction is increasingly used in practice to reduce the time required to provide high-quality images by up to 50%.[1,2] DL image synthesis offers effective 100% series acceleration. In addition, because the synthesized images in our study receive the SNR and spatial resolution of the acquired T1WI and T2WI scans, Syn-STIR can be expected to offer better image quality than is practical with an Acq-STIR.

Previous work on MR imaging sequence-to-sequence translation has been performed[4-10] but generally in subjects without pathology. This study demonstrated excellent performance in a patient cohort with a diverse set of typical spinal pathologies and evaluated key imaging findings commonly assessed with STIR imaging.

Absolute quality ratings could potentially obscure subtle failures and artifacts in the synthetically reconstructed image. Thus, a blinded, side-by-side evaluation was performed to compare the extent of disease and diagnostic confidence as well as to interrogate for evidence of image aberrations. We found that 96% of Syn-STIR sets manifested equal or better extent of disease compared with Acq-STIR, and 88% of cases provided equal or higher

diagnostic confidence with Syn-STIR. Most important, no unexpected image appearances ("hallucinations") or information losses were detected. Although our study had no cases in which the network failed to generate an acceptable Syn-STIR image, the quality of the Syn-STIR image depends on the quality of the input T1 and T2 images. Therefore, if the input images were to have gross artifacts or high noise levels, these could manifest on the Syn-STIR series.

Our study patients were imaged on scanners of differing vendors and field strengths, drawn from a variety of geographically diverse facilities, and encompassed a variety of disease entities, but we acknowledge a risk of inadvertent patient-selection bias or disease-representation bias during the initial gathering of the larger patient cohort.

In this randomized, blinded trial, Syn-STIR demonstrated superior image quality with respect to Acq-STIR. A potential limitation is that the overall qualitative image-quality assessment was a collective summary of perceived metrics against typical expectations and is thus biased by subjective preferences. However, quantitative measures, such as statistical analysis of voxel consistency across STIR data sets, were robust. Future exploration could apply synthetic image generation to additional body parts and other scanning techniques.

CONCLUSIONS

DL-based Syn-STIR MR images, derived from acquired T1WI and T2WI DICOM data sets from multiple centers, scanners, and field strengths, proved statistically interchangeable in diagnostic performance with traditionally acquired STIR and provided superior perceived image quality. Quantitative measures demonstrated consistent results, validating both the high accuracy of the Syn-STIR images and the generalizability of the DL method. This Syn-STIR method offers a promising clinical solution for faster and more comfortable spine MR imaging examinations.

Disclosure forms provided by the authors are available with the full text and PDF of this article at www.ajnr.org.

REFERENCES

1. Bash S, Johnson B, Gibbs W, et al. **Deep learning image processing enables 40% faster spinal MR scans which match or exceed quality of standard of care**. *Clin Neuroradiol* 2022;32:197–203 CrossRef Medline

2. Bash S, Wang L, Airriess C, et al. **Deep learning enables 60% accelerated volumetric brain MRI while preserving quantitative performance: a prospective, multicenter, multireader trial**. *AJNR Am J Neuroradiol* 2021;42:2130–37 CrossRef Medline

3. Zaharchuk G, Gong E, Wintermark M, et al. **Deep learning in neuroradiology**. *AJNR Am J Neuroradiol* 2018;39:1776–84 CrossRef Medline

4. Liu J, Pasumarthi S, Duffy B, et al. **One model to synthesize them all: multi-contrast multi-scale transformer for missing data imputation**. *IEEE Trans Med Imaging* 2023 Mar 4 [Epub ahead of print] CrossRef Medline

5. Li H, Paetzold JC, Sekuboyina A, et al. **DiamondGAN: unified multi-modal generative adversarial networks for MRI sequences synthesis**. In: Shen E, Liu T, Peters, TM, et al. *Medical Image Computing and Computer Assisted Intervention–MICCAI*. Springer International Publishing; 2019:795–803

6. Denck J, Guehring J, Maier A, et al. **Enhanced magnetic resonance image synthesis with contrast-aware generative adversarial networks**. *J Imaging* 2021;7:133 CrossRef Medline

7. Yurt M, Dar SU, Erdem A, et al. **mustGAN: multi-stream Generative Adversarial Networks for MR image synthesis**. *Med Image Anal* 2021;70:101944 CrossRef Medline

8. Dar SU, Yurt M, Karacan L, et al. **Image synthesis in multi-contrast MRI with conditional generative adversarial networks**. *IEEE Trans Med Imaging* 2019;38:2375–88 CrossRef Medline

9. Lee D, Kim J, Moon WJ, et al. **CollaGAN: Collaborative GAN for Missing Image Data Imputation**. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE*, Long Beach, California. June 15–20, 2019:2487–96 CrossRef

10. Haubold J, Demircioglu A, Theysohn JM, et al. **Generating virtual short tau inversion recovery (STIR) images from T1- and T2-weighted images using a conditional generative adversarial network in spine imaging**. *Diagnostics (Basel)* 2021;11:1542 CrossRef Medline

11. Obuchowski NA, Subhas N, Schoenhagen P. **Testing for interchangeability of imaging tests**. *Acad Radiol* 2014;21:1483–89 CrossRef Medline

12. Kingma DP, Ba J. **Adam: A Method for Stochastic Optimization**. *arXiv*. December 22, 2014. http://arxiv.org/abs/1412.6980. Accessed January 15, 2023

13. Khatri G, Pedrosa I, Ananthakrishnan L, et al. **Abbreviated-protocol screening MRI vs. complete-protocol diagnostic MRI for detection of hepatocellular carcinoma in patients with cirrhosis: an equivalence study using LI-RADS v2018**. *J Magn Reson Imaging* 2020;51:415–25 CrossRef Medline

14. Giavarina D. **Understanding Bland Altman analysis**. *Biochem Med (Zagreb)* 2015;25:141–51 CrossRef Medline

15. Altman DG, Bland JM. **Measurement in medicine: the analysis of method comparison studies**. *The Statistician* 1983;32:307–17

16. Henderson AR. **Testing experimental data for univariate normality**. *Clin Chim Acta* 2006;366:112–29 CrossRef Medline

17. Sotgia S, Zinellu A, Pinna GA, et al. **A new general regression-based approach for method comparison studies**. *Clin Chem Lab Med* 2008;46:1046–49 CrossRef Medline

18. Tanenbaum L. **Quality, efficiency, and survival with patient centric imaging**. In: *Proceedings of 7th Annual Snowmass 2023: Hot Topics in Radiology: Advanced Imaging and Artificial Intelligence*. Snowmass Village, Colorado. February 5–10, 2019

19. Meléndez JC, McCrank E. **Anxiety-related reactions associated with magnetic resonance imaging examinations**. *JAMA* 1993;270:745–47 CrossRef Medline

20. Zaitsev M, Maclaren J, Herbst M. **Motion artifacts in MRI: a complex problem with many partial solutions**. *J Magn Reson Imaging* 2015;42:887–901 CrossRef Medline

21. Andre JB, Bresnahan BW, Mossa-Basha M, et al. **Toward quantifying the prevalence, severity, and cost associated with patient motion during clinical MR examinations**. *J Am Coll Radiol* 2015;12:689–95 CrossRef Medline