Radiology

REVIEWS AND COMMENTARY • EDITORIAL

Advancing AI-assisted US Screening for Fatty Liver

Theresa A. Tuthill, PhD

Dr Tuthill is the digital research manager for the US research group at GE HealthCare. Her research interests include quantitative imaging biomarkers and AI strategies for medical imaging. Dr Tuthill is a member of American Institute of Ultrasound in Medicine and serves as a co-chair for the RSNA Quantitative Imaging Biomarkers Alliance Pulse-Echo Quantitative Ultrasound (PEQUS) Committee.

Nonalcoholic fatty liver disease (NAFLD) is often diagnosed in patients originally referred after incidentally noted hepatic steatosis at abdominal imaging (1). Early detection is important, as high liver fat levels (>5% of fat within hepatocytes) are linked to a higher risk of extrahepatic health issues, such as renal and cardiovascular diseases. While the prevalence of NAFLD is estimated at over 25% of the general population, with a number of patients continuing on to develop nonalcoholic steatohepatitis (NASH), patients are usually asymptomatic. This has led to challenges in identifying NAFLD in its early stages, when steatosis remains reversible.

The accepted clinical reference standard for diagnosing NASH remains biopsy with quantitative scoring of fibrosis, steatosis, lobular inflammation, and ballooning. The most accurate noninvasive measurement of steatosis is MRI-based proton density fat fraction (2), but high costs and limited portability hinder its use in screening. Conventional B-mode US, with its low cost and high accessibility, is thus more commonly used to screen for liver steatosis; however, it suffers from interobserver variability. Subjective evaluation is based on signal attenuation, echogenicity (particularly compared with that of the kidney), and appearance of intrahepatic vessels.

To standardize image evaluation, more researchers are relying on the analysis capabilities offered by artificial intelligence (AI) and, more specifically, machine learning (ML). The exponential use of AI in radiology has provided numerous applications of automated interpretation in diagnostic imaging. While certainly not replacing radiologists, AI is viewed as a valuable tool in clinical decision-making and should be embraced (3). Fundamentally, ML is the development of a pattern-identifying model based on extrapolation from training sets that are representative of those in anticipated use. Deep learning (DL) is a subset of ML that uses neural networks to mimic the human learning process. The portability and accessibility of US to less experienced operators puts it at the forefront for AI assistance. Yet, US training sets offer additional challenges to DL modeling due to large image variations from operator and scanner dependencies, as well as patient influences (4).

In this issue of *Radiology*, Vianna et al (5) assessed the classification agreement of radiologists and a DL algorithm in detecting liver steatosis from B-mode US scans. With use of a standard clinical protocol, images were collected from a total of 199 patients within 1 year of a liver biopsy. The pathologic assessment included grading for steatosis on a standard four-point scale (S0 = none, S1 = mild, S2 = moderate, and S3 = severe), which was then used as the reference standard. Selected patients had histologically confirmed NAFLD, NASH, or NASH-related cirrhosis, and the control subset (57 patients) had grade S0 steatosis and no inflammation or fibrosis.

US scan sets of three views per patient were read by six radiologists with varying experience, from junior resident to senior-level subspecialist. Before image interpretation, the six radiologists were trained in standardized assessment of image features such as attenuation, vessel border delineation, and backscatter intensity in relation to the adjacent kidney. A double read paradigm was used with sessions 2 weeks apart to minimize recall bias. Similar to the pathologic assessment, scoring was performed on a four-point ordinal scale, from S0 to S3. In assessing no steatosis from mild or greater steatosis (S0 vs S1 or higher), the intrareader agreement was moderate (κ = 0.45), while the interreader agreement was only fair (κ = 0.34). The corresponding diagnostic performance had areas under the receiver operating characteristic curve (AUCs) ranging from 0.49 to 0.84. This is often considered the more difficult classification, and indeed, other comparisons (S0 or S1 vs S2 or S3 and S2 or lower vs S3) had higher diagnostic performance, with AUCs ranging from 0.57 to 0.76 and 0.52 to 0.81, respectively.

With use of an independent training set along with biopsy scores, a DL algorithm was developed using a convolutional neural network architecture with multiple layers for binary classification. A fivefold cross-validation allowed hyperparameter tuning. The eventual test set was identical to that of the reader evaluation to allow for a direct comparison. The AI diagnostic performance for classifying steatosis grade S0 versus S1 or higher was as good (AUC, 0.85) as that of the top radiologist and had an accuracy of 79%. The AUC for the detection of steatosis was higher than that for grading severity. For the other dichotomized steatosis classes, there was no statistically significant difference between the model and reader performance. Overall,

From GE HealthCare, 500 W Monroe St, Chicago, IL 60661. Received September 11, 2023; revision requested September 14; revision received September 16; accepted September 19. **Address correspondence to** the author (email: *tuthillt@gmail.com*).

Conflicts of interest are listed at the end of this article.

See also the article by Vianna et al in this issue.

the classification accuracies of radiologists and DL with respect to dichotomized steatosis grades were equivalent.

Numerous other studies have shown strong diagnostic capabilities of AI and/or ML methods, though most do not include radiologic reading comparisons and are limited to data sets from a single scanner. The study by Vianna et al used seven different scanners from a range of manufacturers, thus demonstrating robust capabilities. A recent review (6) of liver US AI methods noted key factors that have not been standardized, which included the preprocessing of images and region of interest selection. In the study by Vianna et al (5), cropping was only done to remove screen information; there was no manually selected region of interest.

One of the advantages of DL processes is the creation of a class activation map for each image. Such maps help identify the highly discriminative regions used by the convolutional neural network in classification and can often be used to train readers in identifying visual cues. In the study by Vianna et al, class activation maps were used to understand sources of incorrectly classified images. A few maps exhibited artifacts such as high activation outside the standard US format. This demonstrates the need for human oversight but may also strengthen the case for region of interest placement for improved classification.

One of the challenges that all AI-enabled biomarkers must overcome to be adopted in clinical practice is the interpretability of the algorithms (7). The black-box approach is viable for basic screening, such as detecting mild to moderate steatosis in this case, but key radiomic features used in the algorithm are inaccessible, and thus, biologic rationale is lacking for treatment decisions. In the study by Vianna et al (5), the data set was not large enough to assess potential confounding effects of other NASH-related features such as fibrosis, inflammation, and ballooning.

The biopsy reference standard used in training also limits application to general staging of liver steatosis. For broader use in characterizing and monitoring liver health, a more quantitative approach is needed. A number of initiatives are currently underway to standardize quantitative assessments of visual features. As fat accumulation increases echogenicity, backscatter can be quantified by either the hepatic-renal ratio or with use of the backscatter coefficient, which is independent of machine factors, using a reference phantom method (8). Because the presence of fat also increases the attenuation of the propagating ultrasound beam, calculating the energy loss with use of either temporal or frequency domain approaches has demonstrated good clinical results (9). These methods as well as potential AI and/or ML algorithms should be validated with MRI proton density fat fraction to allow more continuous rather than categorical scoring.

The overall results of this study provide a foundation for using AI and/or ML algorithms on liver B-mode US scans to serve as a screening tool for hepatic steatosis. If these algorithms are used as an aid in assessing radiologic reads, the accuracy in detecting mild to moderate liver steatosis could increase, especially for radiologists with less experience or operating in distant locations relying on telehealth. Combined with simple blood tests, this AI technology would also allow a wider range of clinicians to triage patients deemed to be at high risk for NASH.

Yet, adoption of any AI and/or ML technology requires regulatory approval. In the United States, the current regulatory framework, based on the U.S. Food and Drug Administration's "Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan" (10), encourages development of ML-based diagnostics to incorporate broad ranges of populations in training sets to reduce algorithmic bias. Elimination of factors associated with any race, socioeconomic status, or concomitant diseases would increase the robustness and generalizability of these algorithms. Thus, additional studies, mimicking real-world use, are needed to fully implement AI-integrated approaches to detect early-stage liver steatosis at US.

Disclosures of conflicts of interest: T.A.T. No relevant relationships.

References

1. Rinella ME, Neuschwander-Tetri BA, Siddiqui MS, et al. AASLD practice guidance on the clinical assessment and management of nonalcoholic fatty liver disease. Hepatology 2023;77(5):1797–1835.
2. Yokoo T, Bydder M, Hamilton G, et al. Nonalcoholic fatty liver disease: diagnostic and fat-grading accuracy of low-flip-angle multiecho gradient-recalled-echo MR imaging at 1.5 T. Radiology 2009;251(1):67–76.
3. Yang L, Ene IC, Arabi Belaghi R, Kfoff D, Stein N, Santaguida PL. Stakeholders' perspectives on the future of artificial intelligence in radiology: a scoping review. Eur Radiol 2022;32(3):1477–1495.
4. Akkus Z, Cai J, Boonrod A, et al. A survey of deep-learning applications in ultrasound: artificial intelligence–powered ultrasound for improving clinical workflow. J Am Coll Radiol 2019;16(9 Pt B):1318–1328.
5. Vianna P, Calce SI, Boustros P, et al. Comparison of radiologists and deep learning for US grading of hepatic steatosis. Radiology 2023;309(1):e230659.
6. Alshagathrh FM, Househ MS. Artificial intelligence for detecting and quantifying fatty liver in ultrasound images: a systematic review. Bioengineering (Basel) 2022;9(12):748.
7. Bera K, Braman N, Gupta A, Velcheti V, Madabhushi A. Predicting cancer outcomes with radiomics and artificial intelligence in radiology. Nat Rev Clin Oncol 2022;19(2):132–146.
8. Wear KA, Han A, Rubin JM, et al. US backscatter for liver fat quantification: an AIUM-RSNA QIBA Pulse-Echo Quantitative Ultrasound initiative. Radiology 2022;305(3):526–537.
9. Ferraioli G, Kumar V, Ozturk A, Nam K, de Korte CL, Barr RG. US attenuation for liver fat quantification: an AIUM-RSNA QIBA Pulse-Echo Quantitative Ultrasound initiative. Radiology 2022;302(3):495–506.
10. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. U.S. Food & Drug Administration. Published January 2021. Accessed September 9, 2023.