AJNR

MRI-Based Prediction of Clinical Improvement after Ventricular Shunt Placement for Normal Pressure Hydrocephalus: Development and Evaluation of an Integrated Multisequence Machine Learning Algorithm

Owen P. Leary, Zhusi Zhong, Lulu Bi, Zhicheng Jiao, Yu-Wei Dai, Kevin Ma, Shanzeh Sayied, Daniel Kargilis, Maliha Imami, Lin-Mei Zhao, Xue Feng, Gerald Riccardello, Scott Collins, Konstantina Svokos, Abhay Moghekar, Li Yang, Harrison Bai, Petra M. Klinge and Jerrold L. Boxerman

AJNR Am J Neuroradiol published online 12 June 2024
http://www.ajnr.org/content/early/2024/09/12/ajnr.A8372

Published September 12, 2024 as 10.3174/ajnr.A8372

ORIGINAL RESEARCH
NEURODEGENERATIVE DISORDER IMAGING

MRI-Based Prediction of Clinical Improvement after Ventricular Shunt Placement for Normal Pressure Hydrocephalus: Development and Evaluation of an Integrated Multisequence Machine Learning Algorithm

Owen P. Leary, Zhusi Zhong, Lulu Bi, Zhicheng Jiao, Yu-Wei Dai, Kevin Ma, Shanzeh Sayied, Daniel Kargilis, Maliha Imami, Lin-Mei Zhao, Xue Feng, Gerald Riccardello, Scott Collins, Konstantina Svokos, Abhay Moghekar, Li Yang, Harrison Bai, Petra M. Klinge, and Jerrold L. Boxerman

ABSTRACT

BACKGROUND AND PURPOSE: Symptoms of normal pressure hydrocephalus (NPH) are sometimes refractory to shunt placement, with limited ability to predict improvement for individual patients. We evaluated an MRI-based artificial intelligence method to predict postshunt NPH symptom improvement.

MATERIALS AND METHODS: Patients with NPH who underwent MRI before shunt placement at a single center (2014–2021) were identified. Twelve-month postshunt improvement in mRS, incontinence, gait, and cognition were retrospectively abstracted from clinical documentation. 3D deep residual neural networks were built on skull-stripped T2-weighted and FLAIR images. Predictions based on both sequences were fused by additional network layers. Patients from 2014–2019 were used for parameter optimization, while those from 2020–2021 were used for testing. Models were validated on an external validation data set from a second institution ($n = 33$).

RESULTS: Of 249 patients, $n = 201$ and $n = 185$ were included in the T2-based and FLAIR-based models according to imaging availability. The combination of T2-weighted and FLAIR sequences offered the best performance in mRS and gait improvement predictions relative to models trained on imaging acquired by using only 1 sequence, with area under the receiver operating characteristic (AUROC) values of 0.7395 [0.5765–0.9024] for mRS and 0.8816 [0.8030–0.9602] for gait. For urinary incontinence and cognition, combined model performances on predicting outcomes were similar to FLAIR-only performance, with AUROC values of 0.7874 [0.6845–0.8903] and 0.7230 [0.5600–0.8859].

CONCLUSIONS: Application of a combined algorithm by using both T2-weighted and FLAIR sequences offered the best image-based prediction of postshunt symptom improvement, particularly for gait and overall function in terms of mRS.

ABBREVIATIONS: AI = artificial intelligence; AUROC = area under the receiver operating characteristic curve; CCI = Charlson Comorbidity Index; DESH = disproportionately enlarged subarachnoid space hydrocephalus; iNPH = idiopathic normal pressure hydrocephalus; IQR = interquartile range; ML = machine learning; NPH = normal pressure hydrocephalus; sNPH = secondary normal pressure hydrocephalus

Normal pressure hydrocephalus (NPH) is a progressive neurologic disorder characterized by a diagnostic triad of presenting symptoms, including gait instability, urinary incontinence, and cognitive impairment.[1-3] Etiologically, NPH is understood as a form of communicating hydrocephalus that may result from impaired CSF clearance and re-absorbance in the brain.[3,4] While difficult to diagnose and most frequently idiopathic (iNPH),[5] NPH can also occur secondary to other conditions that influence intracranial CSF dynamics including traumatic brain injury, meningitis, stroke, hemorrhage, or brain tumor (secondary normal pressure hydrocephalus [sNPH]).[6] The mainstay of treatment for NPH is surgical placement of a ventricular shunt, which drains CSF from the cerebral ventricles to the peritoneal (ventriculoperitoneal) or less commonly the pleural (ventriculopleural) space.[2,7,8]

Received January 2, 2024; accepted after revision May 16.

From the Departments of Neurosurgery (O.P.L., K.M., S.S., K.S., P.M.K.), and Diagnostic Imaging (Z.Z., L.B., Z.J., G.R., S.C., J.L.B.), Brown University Warren Alpert Medical School, Providence, Rhode Island; School of Electronic Engineering (Z.Z.), Xidian University, Xi'an, China; Department of Neurology (Y.-W.D., L.Y.), The Second Xiangya Hospital, Central South University, Hunan, China; Columbia University Vagelos College of Physicians and Surgeons (K.M.), New York, New York; Departments of Radiology (D.K., M.I., L.-M.Z., H.B.), and Neurology (A.M.), Johns Hopkins University School of Medicine, Baltimore, Maryland; Carina Medical (X.F.), Lexington, Kentucky; and Department of Biomedical Engineering (X.F.), University of Virginia, Charlottesville, Virginia.

Owen P. Leary and Zhusi Zhong contributed equally to this work.

Please address correspondence to Owen P. Leary, BS, Warren Alpert Medical School of Brown University, Department of Neurosurgery, Rhode Island Hospital, APC Building, Floor 6, 593 Eddy St, Providence RI 02903; e-mail: owen_leary@brown.edu

http://dx.doi.org/10.3174/ajnr.A8372

SUMMARY

PREVIOUS LITERATURE: Prior literature has highlighted multiple imaging-based metrics on CT and MRI (eg, disproportionately enlarged subarachnoid space hydrocephalus score) that can be useful in diagnosis and outcome prediction of normal pressure hydrocephalus, a neurologic disorder defined by a triad of symptoms including gait abnormality, urinary dysfunction, and cognitive impairment. Prediction of improvement after shunt placement surgery remains a challenge, impeding optimal clinical decision-making. We applied artificial intelligence approaches to automatically extract radiomic features from multiple MRI sequences (T2-weighted and FLAIR) acquired preoperatively from a single-center cohort of patients with normal pressure hydrocephalus and evaluated the performance of models trained to predict function and symptom improvements postoperatively.

KEY FINDINGS: Artificial intelligence algorithms leveraging combined imaging features from preoperative T2-weighted and FLAIR sequences were generally more predictive of postoperative shunt outcome in normal pressure hydrocephalus than models built by using 1 of these sequences. Models performed best for prediction of gait and incontinence improvement, and slightly worse for overall function and cognition.

KNOWLEDGE ADVANCEMENT: We demonstrate implementation of an analytic pipeline for automated radiomic feature extraction from multiple MRI sequences and layered integration of those data to optimize and evaluate prognostic models for normal pressure hydrocephalus, a complex and poorly understood neurologic disorder with defined symptoms and treatment options.

While considered an effective treatment strategy for many patients, symptoms of NPH are sometimes refractory to shunt surgery, with 15%–30% of patients experiencing little improvement across symptom domains.[2] Furthermore, traditional methods of simulating CSF drainage, such as high-volume lumbar tap test or CSF dynamic testing, have failed to predict reliably which individuals are most likely to benefit from shunting, with negative predictive values of ≤50%.[9-12] While some evidence supports the utility of comorbidity status and other clinical variables in predicting outcome of shunt surgery, models based on comorbidity status alone do not markedly advance clinical decision-making capacity.[13-15] Furthermore, imaging biomarkers derived from single-technique imaging (eg, T2-weighted MRI alone) have failed to significantly improve the overall ability to predict patient outcome.[3,16,17]

With artificial intelligence (AI) models increasingly applied to prognostication, there is new opportunity to leverage information drawn from multiple MRI sequences available upon baseline neuroimaging in NPH. While of limited utility in isolation, MRI-derived markers combined into a unified model could advance NPH diagnosis and decision-making.[11,16-18] In the present study, we hypothesized that AI driven models trained on MRI sequences sensitive to both structural and CSF distribution parameters may predict clinical benefit after shunting for NPH, and that models using multisequence information may be the best performing.

MATERIALS AND METHODS

Primary Clinical Data Set

The primary data set included 249 consecutive patients who underwent ventricular shunting for clinical diagnosis of NPH who were retrospectively identified from a single institutional cohort (Rhode Island Hospital, Providence, Rhode Island). Two hundred twelve patients (85.1%) were diagnosed with iNPH and 37 patients (14.9%) had sNPH. All were treated by 1 of 2 neurosurgeons (P.M.K. and K.S.) from 2015–2021. Clinical data were abstracted from recorded clinical documentation by trained research staff. Baseline characteristics including comorbidity status as quantified in the Charlson Comorbidity Index (CCI) were also obtained.[14] Postoperatively, 226 patients (90.8%) returned to clinic at 3-month follow-up and 175 (70.3%) returned for 12-month follow-up. For outcome analysis, 12-month follow-up was preferentially selected whenever available, however if 12-month follow-up was not available, 3-month follow-up was utilized for evaluation of the symptom improvement end points.

MRI sequences of interest included T2-weighted and FLAIR, selected based on availability as part of routine diagnostic MRI acquired during clinical work-up at our center and the hypothesis that these sequences may provide complementary information about underlying disease features (Fig 1). A total of 201 patients (80.7%) had preoperative imaging (ie, T2-weighted

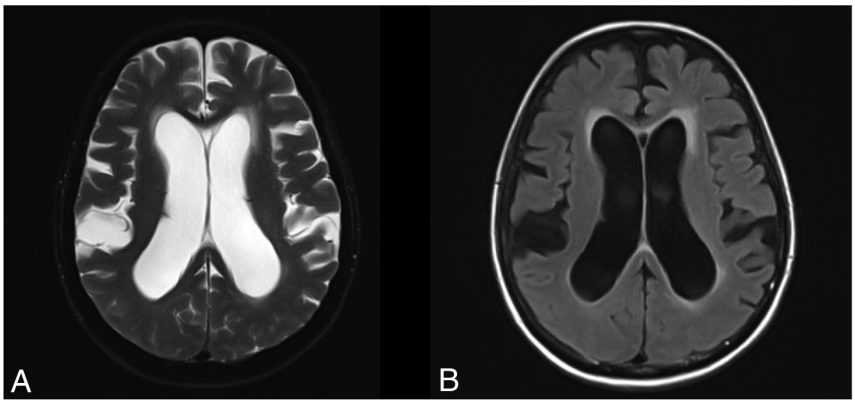

FIG 1. Exemplary axial images from T2-weighted (*A*) and FLAIR (*B*) sequences of an included patient.

or FLAIR) available for inclusion. Due to lack of imaging and/or follow-up data, 48 of 249 patients were excluded from the FLAIR-based analysis, and 64 of 249 patients were excluded from the T2 and T2+FLAIR analyses. Most scans were acquired at a single academic center by using a uniform image acquisition protocol, by using either a 1.5T or 3T magnet, though a minority of cases were captured at outside centers. The acquisition protocols on both scanners were the same except that the slice thickness was 5 mm for the 1.5T and 4 mm for 3T, which is accounted for by preprocessing methods described subsequently. No notable changes in NPH referral processes occurred during the study inclusion period (2015–2021). Accordingly, baseline MRI data from included patients was separated into training and testing sets based on date (approximate 7:3 ratio).

Gait and urinary incontinence scores were estimated by using previously reported 8- and 6-point symptom scales, respectively.[1,2] For the gait scale: 1 = normal; 2 = slight disturbance of tandem walk and turning; 3 = wide-based gait with sway; 4 = tendency to fall, with foot corrections; 5 = walking with cane; 6 = bimanual support; 7 = aided by another person; and 8 = wheelchair-bound. For the incontinence scale: 1 = normal; 2 = urgency without incontinence; 3 = infrequent incontinence; 4 = frequent incontinence; 5 = complete bladder incontinence; and 6 = bladder/bowel incontinence. Cognitive impairment was evaluated as present or absent based on patient report and clinician impression. In addition, global functional disability was estimated retrospectively by using mRS.[19] All of these metrics were estimated based on descriptive reports of patient-reported symptoms and examination findings in the history and examination sections of baseline and follow-up clinical notes by the 2 treating surgeons; whenever ambiguity existed between 2 levels of a given score, the lesser of the 2 was selected. Improvements in gait, incontinence, and mRS were defined by ≥1 point improvement and improvement in cognition was defined by the patient's and clinician's impression. Binarized clinical end points (improved versus not improved) in each symptom domain and mRS were recorded at 3-months and 12-months postoperatively. Institutional review board authorization was obtained for chart and imaging review (#1345067).

External Validation Data Set

To assess generalizability of findings to external NPH cohorts, we obtained an external validation data set comprising 33 patients with shunted NPH with complete follow-up data treated at a second institution (Johns Hopkins Hospital, Baltimore, Maryland) within the same period. The external validation data set comprised a nonconsecutive convenience sample of patients with sufficient follow-up and imaging data from the same treatment period. Imaging sequences and acquisition protocols were generally the same as described above (1.5T or 3T magnet with variable slice thickness, but preprocessed uniformly before subsequent feature extraction).

Preprocessing

All the FLAIR and T2-weighted images were preprocessed according to the pipeline demonstrated in Fig 2. The OncoAI algorithm first registers FLAIR and T2-weighted sequences to a standard template, performs brain extraction, concatenates all sequences, deploys the pretrained segmentation to obtain the final label map, warps labels back to the original space, and performs limited, stereotyped postprocessing.[20] For each study with multiple sequences, the algorithm first parses all series, skips localizer and calibration images, coregisters all series to a standard template, and resamples to $1 \times 1 \times 1$ mm^3 voxel size. Then, the algorithm finds the series that has the largest coverage and runs an AI-based brain segmentation model to extract the brain and warps the brain mask back to each series. The brain space was parcellated from the skull ("skull stripping") and centered within the volume, followed by automated segmentation of intracranial tissue structures (Fig 2). MRI images were resized to $256 \times 256 \times 64$ pixels for the deep neural network. Range scaling normalization was then applied based on intensity value distribution.

Deep Feature Learning

3D deep residual neural network models were separately built on T2-weighted and FLAIR images for each binary clinical improvement end point (ie, "improved" versus "not improved" in gait, incontinence, cognition, and mRS). We used 3D ResNet-50 as the single technique deep feature extraction model (Fig 3). Each block of ResNet-50 is a combination of 3 deep layers with different convolution parameters, in which every 2 residual layers are inserted within this 3-layer bottleneck block, including a total of 50 3D residual layers. After an average pooling layer, the 2048-dimensional image features extracted by ResNet-50 were used to predict classification probability, constructing a classifier with 3 linear layers with ReLU and Drop-out. The numbers of hidden nodes in linear layers were 2048, 2048, and 1024. This pipeline was used for training both the T2- and FLAIR-based models. By contrast, the models trained on both FLAIR and T2-weighted sequences were fused by additional network layers to obtain combined results accounting for both sequences (Fig 4). The features of FLAIR and T2-weighted sequences were extracted by 2 independent deep residual networks, and the 2 2048-dimensional features were fused by cascading and passing the classifier to obtain the multitechnique classification probability.

Clinical Improvement Prediction

We use the cross-entropy loss function during model training. The classification probability estimates the soft target by the SoftMax function, and cross-entropy loss calculates the loss between the soft target and the ground-truth label to learn model parameters:

$$L(y,z) = \sum_{i \div 0}^{M} -y_i \log\left(\frac{z_i}{\sum_j \exp(z_j)}\right)$$

In this function, M is the total number of classes (eg, M = 2 when for binary outcome prediction). The variable y_i is a vector representing the ground-truth label of the training set as 1 and all other elements as 0; z_j is the logit that is the output of the last layer for the j-th class of the model. The weight of the model is updated via adaptive moment estimation with weight decay, in which the optimizer calculates the adaptive learning rates of every parameter. The learning rate and weight decay coefficient were

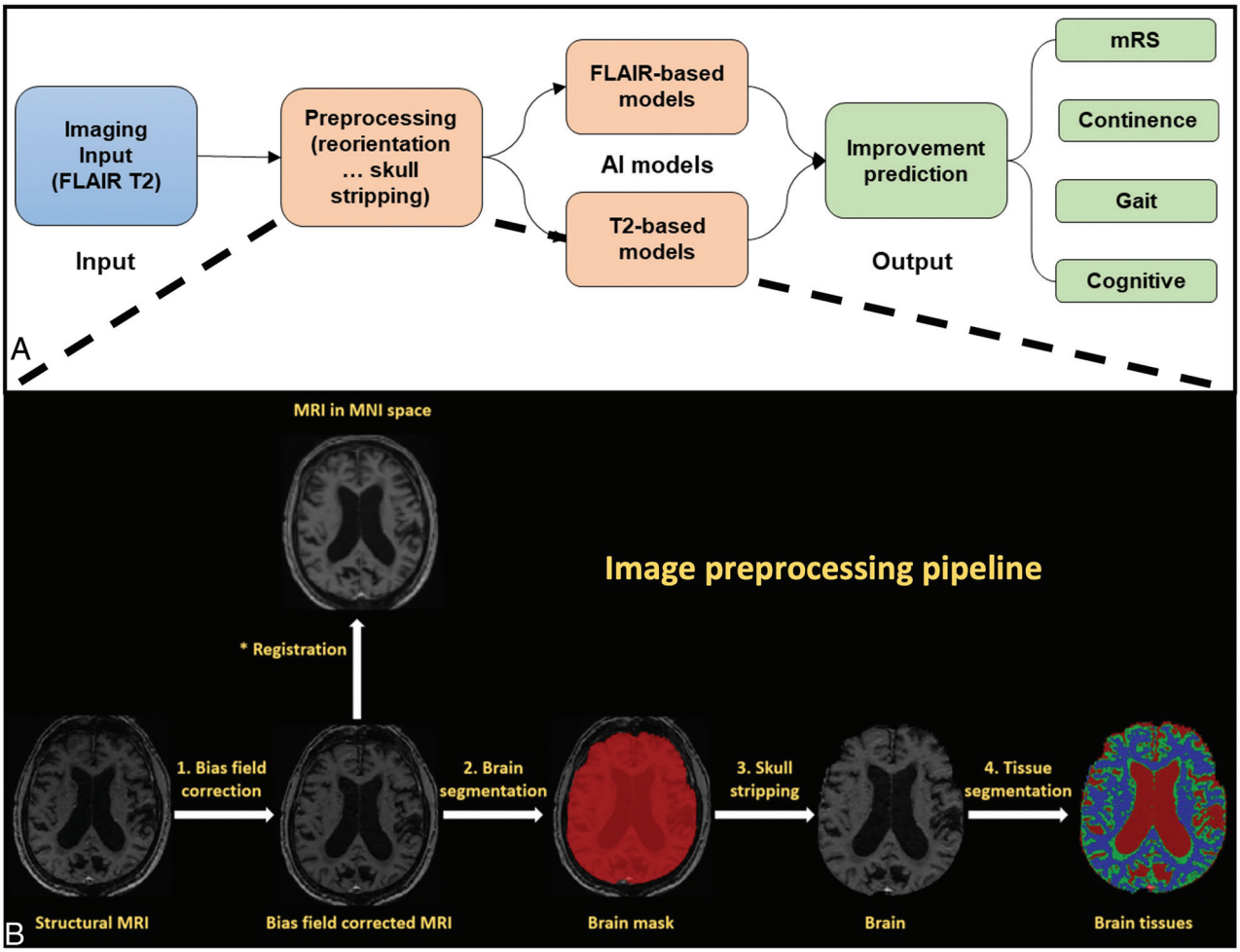

FIG 2. Our AI pipeline for training an outcome prediction model on shunted NPH baseline MRI data set (*A*), including a detailed review of image pre-processing (*B*).

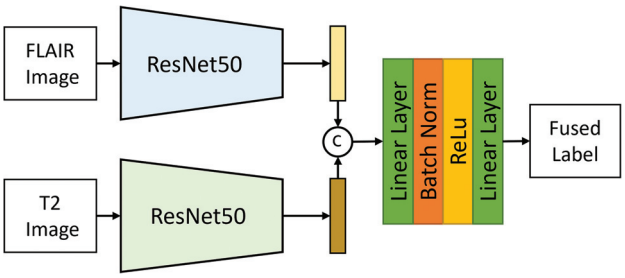

FIG 3. Structure of ResNet50 network.

set to 0.001. We ran each method for 60 epochs and collected the highest average accuracy per run. The fused (T2+FLAIR) model was then initialized with a model trained on a single technique.

Statistical Analysis and Reporting

Overall model performance for T2-weighted alone, FLAIR alone, and T2+FLAIR models are reported in terms of area under the receiver operating characteristic (AUROC), a metric that accounts for the balance between true-positive rate and false-positive rate across decision thresholds. AUROC values were also computed for models trained only on age, sex, and CCI. For training data, we applied class-balance resampling, with equal numbers of non-improved and improved patients to reduce class imbalance. After assessing performance on the primary data set, all models were tested separately on the validation data set. AUROC values are reported with 95% CI.

RESULTS

Patient Cohorts

Of 249 patients, 129 (51.8%) were men, 232 (93.2%) were white, and 212 (85.1%) were diagnosed with iNPH (Table 1).[6] The distal terminus of the shunt catheter was placed intraperitoneally in 244 of 249 (98.0%). All patients presented with gait instability with a median estimated gait score of 4 of 8 (interquartile range [IQR] 4–6), 196 (78.7%) presented with urinary incontinence with a median estimated incontinence score of 3 of 6 (IQR 2–4), and 217 (87.1%) had subjective cognitive impairment (Table 2).[1,2] The median CCI of the population was 6 (IQR 5–7). Of those who returned for 12-month postoperative follow-up, the greatest overall proportional improvements were observed in gait (70.3%), followed by incontinence (68.0%), overall functional disability (56.4%), and cognition (46.9%). The external validation data set comprising 33 patients with shunted NPH from a second

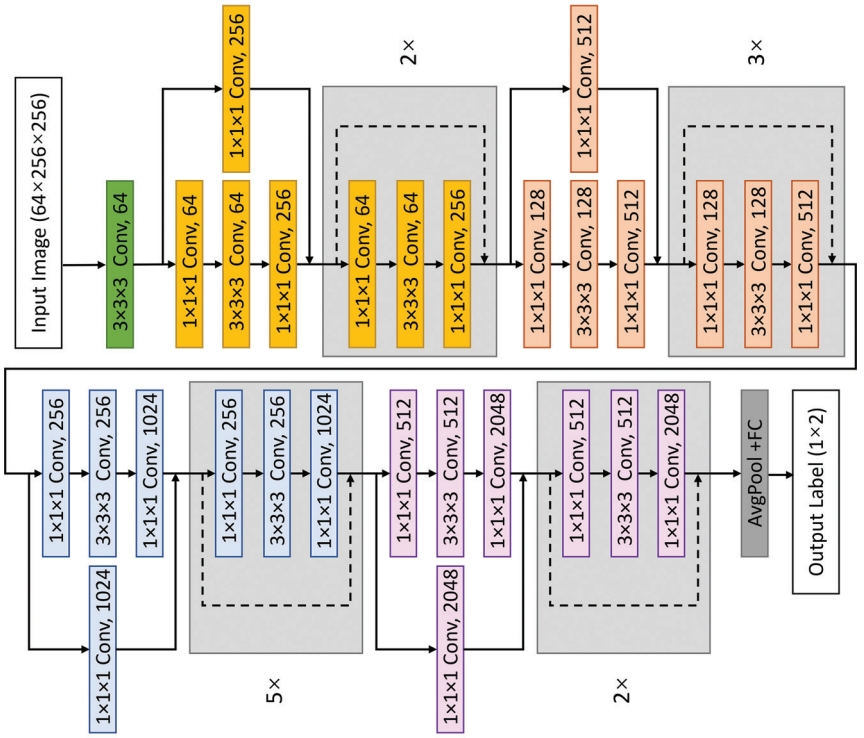

FIG 4. Fusion prediction model with ResNet50 models for integrating information from multiple imaging sequences.

Table 1: Patient demographics and characteristics

Patient Characteristic Mean (± SD) or Median (IQR)	Training/Testing (Institution 1, n = 249)	Validation (Institution 2, n = 33)
Age	74.2 (± 7.5) years	72.1 (± 7.7) years
Sex		
Men	129 (51.8%)	19 (57.6%)
Women	120 (48.2%)	14 (42.4%)
Race		
White	232 (93.2%)	28 (84.8%)
Black/African American	7 (2.8%)	4 (12.1%)
Other/unknown	10 (4.0%)	1 (3.0%)
BMI	28.5 (± 5.8)	29.1 (±4.6)
CCI	6 (5–7)	4 (3–6)
Shunt laterality		
Right	239 (95.9%)	33 (100%)
Left	10 (4.1%)	0 (0%)
Catheter terminus placement		
Ventriculoperitoneal	244 (98.0%)	33 (100%)
Ventriculopleural	5 (2.0%)	0
NPH classification		
iNPH	212 (85.1%)	32 (97.0%)
sNPH	37 (14.9%)	1 (3.0%)
Baseline mRS		
0– No symptoms	0	0
1– No significant disability	1 (0.4%)	0
2– Slight disability	107 (43.0%)	8 (24.2%)
3– Moderate disability	83 (33.3%)	8 (24.2%)
4– Moderate-severe disability	41 (16.5%)	15 (45.5%)
5– Severe disability	3 (1.2%)	1 (3.0%)
Unable to determine	14 (5.6%)	0
8-Point gait score	4 (4–6)	5 (4–6)
6-Point incontinence score	3 (2–4)	3 (2–4)

Note:—BMI indicates body mass index; SD, standard deviation.

institution overall had similar demographic, symptom, and postoperative improvement distributions (Table 1 and 2).

Machine Learning Models

When trained by using T2-weighted imaging alone, the machine learning (ML) model achieved AUROC of 0.5125 [0.3392–0.6858] for postoperative improvement in mRS, 0.6994 [0.5901–0.8087] for gait, 0.7304 [0.6253–0.8355] for urinary incontinence, and 0.4479 [0.2763–0.6195] for cognitive impairment (Table 3). By contrast, models based on FLAIR imaging were generally better performing, yielding AUROC values of 0.6723 [0.4968–0.8478], 0.7195 [0.6046–0.8344], 0.7929 [0.6912–0.8947], and 0.7175 [0.5533–0.8816], respectively. The combination of T2-weighted and FLAIR sequences offered the best performance in mRS and gait improvement predictions relative to models trained on single-sequence imaging, with AUROC values increasing to 0.7395 [0.5765–0.9024] and 0.8816 [0.8030–0.9602]. Performances when predicting the other outcomes were similar to FLAIR-only models (Fig 5). FLAIR-weighted and combined imaging-trained models generally performed better than models trained only on age, sex, and CCI (Table 3). When applied to the external validation data set, T2+FLAIR models were again the best performing for mRS and gait improvements, but similar to single-sequence models for urinary incontinence and cognition (Table 4).

DISCUSSION

Predicting success of shunt surgery in NPH has been historically challenging. Clinical examination and detailed history taking to identify the NPH symptom triad aids in the diagnosis and selection of surgical candidates, however overall prognostic accuracy rarely exceeds 80%–85% mainly due to poor negative predictive values of objective CSF dynamic testing methods.[2,3,10,21] Being able to predict the likelihood of improvement of each symptom domain with greater confidence could boost patient-reported quality of life

Table 2: Presenting symptoms and postoperative improvement; denominators for improvement calculations are patients who had each symptom at baseline and attended follow-up

Triad of NPH Symptoms and Functional Disability	Present at Baseline (Pre-Op)	Improved at 3 Months Post-Op	Improved at 12 Months Post-Op
Training & testing cohort	249/249 (100%)	226/249 (90.8%)	175/249 (70.3%)
Gait impairment	249 (100%)	165/226 (73.0%)	123/175 (70.3%)
Urinary incontinence	196 (78.7%)	76/127 (59.8%)	66/97 (68.0%)
Cognitive impairment	217 (87.1%)	97/208 (46.6%)	75/160 (46.9%)
mRS	—	105/179 (58.7%)	84/149 (56.4%)
Validation cohort	33/33 (100%)	33/33 (100%)	33/33 (100%)
Gait impairment	33 (100%)	29/33 (87.9%)	22/33 (66.7%)
Urinary incontinence	21 (63.6%)	4/33 (12.1%)	3/33 (9.1%)
Cognitive impairment	31 (93.9%)	14/33 (42.4%)	12/33 (36.4%)
mRS	—	7/33 (21.2%)	6/33 (18.2%)

Table 3: Model performances on primary institution data set

Model Input	mRS Improvement	Gait Improvement	Urinary Improvement	Cognitive Improvement
Age, sex, CCI	0.6255 [0.5089–0.7421]	0.5413 [0.4205–0.6621]	0.7135 [0.5306–0.8964]	0.6562 [0.4918–0.8207]
T2 only	0.5125 [0.3392–0.6858]	0.6994 [0.5901–0.8087]	0.7304 [0.6253–0.8355]	0.4479 [0.2763–0.6195]
FLAIR only	0.6723 [0.4968–0.8478]	0.7195 [0.6046–0.8344]	0.7929 [0.6912–0.8947]	0.7175 [0.5533–0.8816]
T2+FLAIR	0.7395 [0.5765–0.9024]	0.8816 [0.8030–0.9602]	0.7874 [0.6845–0.8903]	0.7230 [0.5600–0.8859]

Note:—AUROC values represent the results of pretrained models when evaluated on testing data set and 95% CI.

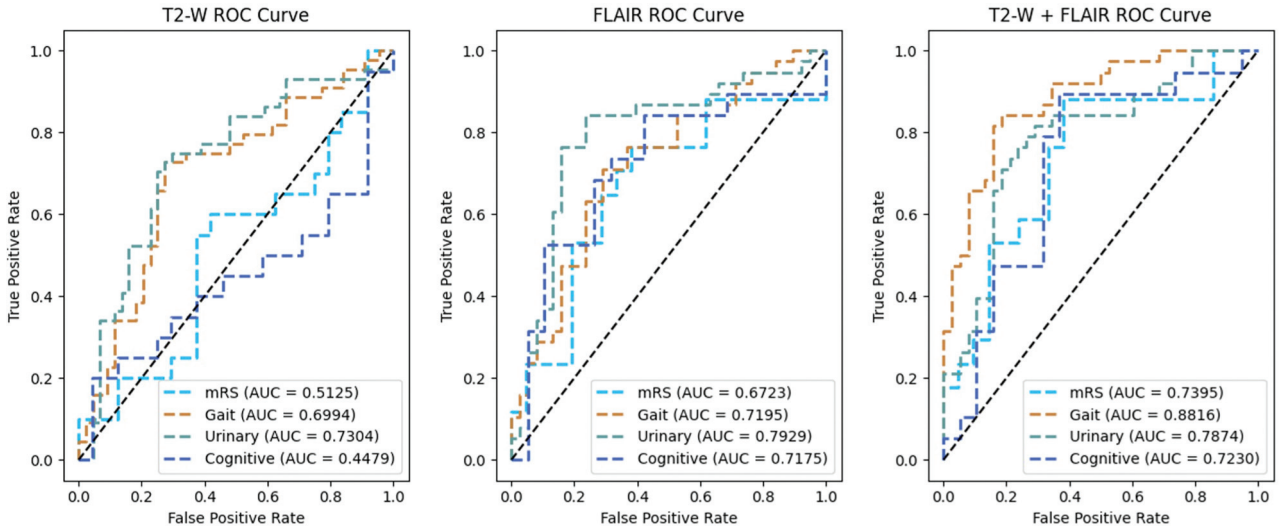

FIG 5. Receiver operating characteristic (ROC) curves demonstrating comparison between models trained on T2-weighted, FLAIR, and both sequences for each of 4 clinical improvement end points on primary institution testing data set.

after shunt surgery and aid more individualized decision-making.[22] In this study, we applied a novel AI-driven approach to outcome prediction by using 2 baseline MRI sequences in patients with shunted NPH. Our methods enabled unbiased feature selection from input imaging (Fig 2) and optimization of outcome prediction across a broader space of potential features relative to standard statistical outcome prediction methods (eg, regression). Our sizable primary study cohort included both patients with iNPH and sNPH diagnosed clinically (Table 1), not dependent on lumbar tap test results.[21] We observed postoperative improvement rates across the NPH triad similar to the reported literature (Table 2).[8,23] In this setting, we hypothesized that MRI-based predictive models achieve optimized performance when combining T2-weighted and FLAIR MRI sequences, each containing clinically useful features of the underlying disease, though this complementarity may pertain to some symptom domains but not others.

Our findings confirm the hypothesis, demonstrating proof-of-concept that ML models trained on both FLAIR and T2-weighted sequences showed improved prediction of treatment success (Table 3). These results confirm the complementary "power" of combining features from multiple sequences into a single model: while overall performance in terms of AUROC was generally less impressive for T2-weighted imaging alone, values increased substantially across mRS and gait outcome domains for the combined (T2+FLAIR) model, and for urinary and cognitive improvement when applying either the FLAIR-only or combined models (Table 3). The optimal model performances for predicting improvement in gait (AUROC = 0.88 [0.80–0.96]), cognition (AUROC = 0.72 [0.56–0.89]), and overall function (AUROC =

Table 4: Model performances on external validation data set

Model Input	mRS Improvement	Gait Improvement	Urinary Improvement	Cognitive Improvement
T2 only	0.5513 [0.2874–0.8152]	0.6167 [0.2391–0.9942]	0.7816 [0.4609–1]	0.5732 [0.4326–0.7139]
FLAIR only	0.7989 [0.6909–0.9068]	0.4758 [0.0517–0.8999]	0.6724 [0.3645–0.9803]	0.6364 [0.5024–0.7703]
T2+FLAIR	0.8291 [0.7271–0.9311]	0.7333 [0.4263–1]	0.7586 [0.4286–1]	0.6310 [0.4326–0.8293]

Note:—AUROC values represent the results of the same pretrained models when evaluated on external validation data set and 95% CI.

0.74 [0.57–0.90]) were observed when applying the T2+FLAIR model, while the maximum performance for predicting urinary symptom improvement was seen with the FLAIR-only model 0.79 [0.69–0.89]. While the T2-weighted sequence offers more detailed information about CSF and subarachnoid space volumes, FLAIR provides information about transependymal flow and white matter change associated with parenchymal volume loss. Though it is possible automatically extracted radiomic features associated with these specific differences between the 2 sequences led to the observed performance differences across symptom domains between T2, FLAIR, and T2+FLAIR models, the methods do not enable us to readily interpret the extracted features to confirm or refute this theory.

Though our methods and results do not define or focus on any individual radiomic feature explicitly, our approach is supported by prior studies that have used more traditional statistical approaches to investigate the value of several structural and CSF space-related imaging features for diagnosis and/or prognosis of NPH.[11,17,18,24,25] MRI-based disproportionately enlarged subarachnoid space hydrocephalus (DESH) score has been found to be statistically disparate between patients who improve versus those who do not across some clinical symptom metrics.[16,24] Furthermore, these findings build on recent studies by Shao et al[26] and others[27,28] who applied ML learning approaches to capture features of the ventricular system and differentiate patients with NPH from non-NPH control individuals, and Tsou et al[29] who demonstrated a convolutional neural network approach to measuring aqueductal CSF flow from phase-contrast MRI. In our study, we used an automated radiomic feature extraction and outcome prediction pipeline that was at least as predictive of symptom improvement across multiple domains as these previously reported approaches. Our approach also did not require radiologic interpretation or measurement of multiple metrics as the DESH score requires.

Additionally, we define an analytic framework for comparing single-sequence versus multisequence models for outcome prediction. While model performance improved by combining T2 and FLAIR when predicting mRS and gait improvements, performance was similar between T2+FLAIR and FLAIR-only trained-models when predicting urinary and cognitive symptom improvement. Taken together, these findings suggest that T2-weighted imaging may not provide additional features useful for predicting these outcomes, while features available on T2-weighted sequences are more sensitive to CSF space and cerebral blood flow characteristics thought to impact gait-related function.[30] Importantly, overall function as evaluated by mRS is more dependent on gait and less dependent on the cognition and urinary incontinence, given the inclusion of mobility-related function explicitly in the score.[8,31]

Further work is needed to understand why some imaging studies, and some specifically identified features within them, may hold more prognostic relevance than others, given overall limited knowledge of the underlying disease mechanisms of NPH. It is our hope that this study will lay the groundwork for a more sophisticated future predictive model which integrates all available data of known prognostic importance. We believe this is among the most comprehensive imaging-based demonstrations of NPH shunt outcome prediction to date, benefitting from an unbiased approach not tethered to any specific hypothesis about individually selected features, as prior studies largely have been.

Comparison with External Validation Data Set

To improve the rigor of our analysis and assess generalizability to external NPH cohorts, we applied our models to an independent NPH population from a second institution. In validation analyses, high performances for predicting mRS, gait, and urinary symptom improvements were maintained, but we observed a drop in performance when predicting cognitive improvement (Table 4). Surprisingly, predictions of improvement in overall function (mRS) with either the FLAIR-only or T2+FLAIR models were even better in the validation cohort, with AUROC values of 0.80 [0.69–0.91] and 0.83 [0.73–0.93], respectively.

The external validation data set was different from the primary data set in 2 important respects: the lower proportion of sNPH patients and the lower rate of improvement across symptom domains. While retrospectively estimated clinical outcomes related to gait and urinary incontinence may be more readily evaluated in the absence of detailed prospective study-specific assessments, cognitive improvement is more difficult to objectively assess in this manner. This may explain the poor performance of the cognitive improvement prediction model on the validation cohort. Furthermore, the confidence intervals of the estimated AUROC value estimates were quite broad, likely because of the small sample size (Table 4, $n = 33$). Confidence intervals were narrower for model evaluation by using the primary data set testing split (Table 3, $n = 55$–60, depending on the model). Accordingly, while limited statistical conclusions can be drawn from comparing model performances without a much larger sample size, the overall performance of the generated models on both the primary testing and external validation data sets provides conceptual evidence that ML models trained on multisequence imaging data may comprise a generalizable approach to improved clinical outcome prediction in NPH. Inclusion of both iNPH and sNPH subpopulations within our cohorts further lends generalizability, because patients with sNPH are often excluded from NPH outcome studies and might be expected to present with more heterogeneous imaging features on MRI. Future work might also seek to leverage additional layers of patient data, such as clinical biomarker data obtained from CSF, which has also recently been shown to offer clinically useful outcome prediction performance.[32,33]

Comparison with Other Models

Within our cohort, models trained on demographic data and CCI without imaging had poor predictive value, underscoring the added value of neuroimaging features that cannot be captured by clinical data alone. Several studies examining correlations between comorbidity data and outcome in NPH have also documented poor prediction based on comorbidity metrics alone.[13,34,35] The limitations of existing predictive tools are particularly apparent when NPH is present alongside Parkinson and/or Alzheimer diseases.[36,37] Multisequence imaging models such as the T2+FLAIR model presented in this study could be further developed to incorporate T1-weighted imaging more sensitive to cognitive outcome,[37] as well as other sequences (eg, diffusion-weighted imaging). Future research is needed to determine the practical value of incorporating these additional sequences not investigated in the present study, but it is certainly plausible that incorporating T1- or DWI-weighted sequences could improve model performance in outcome prediction domains where our T2+FLAIR model is weakest (eg, cognition). Furthermore, outcome prediction models might endeavor to capture more detailed systemic and neurodegenerative comorbidity information than previously published studies of comorbidity burden and NPH outcome,[13-15] given the possibility that such clinical data could also further enhance the performance of models based on imaging alone.

Limitations

While we present the first fully integrated AI model trained on multiple MRI sequences, we selected T2-weighted and FLAIR imaging based on availability through routine clinical practice. Not considered were other sequences such as T1 or DWI sequences that have previously been shown to be of potential importance in NPH because of the retrospective methods of this study.[18,37] While imaging acquisition was generally uniform over the study inclusion period, given that most scans were acquired at a single academic medical center under the same protocol, the minority of patients whose imaging originated from outside centers were not excluded on that basis alone. Accordingly, some heterogeneity in MR acquisition parameters likely exists, though the preprocessing pipeline was intended to partially mitigate any such effect.

We retrospectively abstracted clinical data including gait and continence scores, improvement in symptoms, and mRS from the treating surgeon's clinical documentation, and a portion of patients were lost to follow-up. Given nearly 30% loss to follow-up rate at 12-months postop (versus only 10% at 3 months), the 3-month follow-up was utilized for retrospective outcome assessment for a subset of cases (approximately 20%). While standardized scales were used to estimate impairment scores in the gait and incontinence domains, cognitive impairment was assessed less objectively. Future studies would benefit from standardized implementation of quantitative cognition metrics such as Mini-Mental State Examination.

Finally, while we hypothesized that some of the intuitively observable differences between T2 and FLAIR sequences (eg, transependymal flow on FLAIR) might give rise to their additive value in a combined model, the AI methods utilized do not give us ready access to the automatically identified radiomic features used in predictive modeling, hence this mechanistic aspect of the hypothesis remains unanswered by the present study.

CONCLUSIONS

AI algorithms leveraging combined imaging features from preoperative T2-weighted and FLAIR sequences were generally more predictive of postoperative shunt outcome in NPH than models built by using 1 of these sequences. Models performed best for prediction of gait and incontinence improvement, and slightly worse for predicting improvement in mRS and cognition.

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

REFERENCES

1. Hellström P, Klinge P, Tans J, et al. **A new scale for assessment of severity and outcome in iNPH**. *Acta Neurol Scand* 2012;126:229–37 CrossRef Medline
2. Klinge P, Hellström P, Tans J; European iNPH Multicentre Study Group, et al. **One-year outcome in the European multicentre study on iNPH**. *Acta Neurol Scand* 2012;126:145–53 CrossRef Medline
3. Relkin N, Marmarou A, Klinge P, et al. **Diagnosing idiopathic normal-pressure hydrocephalus.** *Neurosurgery* 2005;57:S4–16 CrossRef Medline
4. Bradley WG Jr. **CSF flow in the brain in the context of normal pressure hydrocephalus.** *AJNR AJNR Am J Neuroradiol* 2015;36:831–38 CrossRef Medline
5. Rigamonti D, Yasar S, Vivas-Buitrago T, et al. **Letter to our colleagues family practitioners, geriatricians, and radiologists to increase awareness regarding idiopathic normal pressure hydrocephalus.** *World Neurosurg* 2023;181:e291–93 CrossRef Medline
6. Daou B, Klinge P, Tjoumakaris S, et al. **Revisiting secondary normal pressure hydrocephalus: does it exist? A review.** *Neurosurg Focus* 2016;41:E6 CrossRef Medline
7. Alvi MA, Brown D, Yolcu Y, et al. **Prevalence and trends in management of idiopathic normal pressure hydrocephalus in the United States: insights from the National Inpatient Sample.** *World Neurosurg* 2021;145:e38–52 CrossRef Medline
8. Giordan E, Palandri G, Lanzino G, et al. **Outcomes and complications of different surgical treatments for idiopathic normal pressure hydrocephalus: a systematic review and meta-analysis.** *J Neurosurg* 2018;131:1024–36 Medline
9. El Ahmadieh TY, Wu EM, Kafka B, et al. **Lumbar drain trial outcomes of normal pressure hydrocephalus: a single-center experience of 254 patients.** *J Neurosurg* 2019;132:306–12 CrossRef Medline
10. Ishikawa M, Hashimoto M, Mori E, et al. **The value of the cerebrospinal fluid tap test for predicting shunt effectiveness in idiopathic normal pressure hydrocephalus.** *Fluids Barriers CNS* 2012;9:1 CrossRef Medline
11. Lotan E, Damadian BE, Rusinek H, et al. **Quantitative imaging features predict spinal tap response in normal pressure hydrocephalus.** *Neuroradiology* 2022;64:473–81 CrossRef Medline
12. Rydja J, Eleftheriou A, Lundin F. **Evaluating the cerebrospinal fluid tap test with the Hellström iNPH scale for patients with idiopathic normal pressure hydrocephalus**. *Fluids Barriers CNS* 2021;18:18 CrossRef Medline
13. Israelsson H, Larsson J, Eklund A, et al. **Risk factors, comorbidities, quality of life, and complications after surgery in idiopathic normal pressure hydrocephalus: review of the INPH-CRasH study.** *Neurosurg Focus* 2020;49:E8 CrossRef Medline
14. Klinge PM, Ma KL, Leary OP, et al. **Charlson comorbidity index applied to shunted idiopathic normal pressure hydrocephalus.** *Sci Rep* 2023;13:5111 CrossRef Medline

15. Pyykkö OT, Nerg O, Niskasaari HM, et al. **Incidence, comorbidities, and mortality in idiopathic normal pressure hydrocephalus**. *World Neurosurg* 2018;112:e624–31 CrossRef Medline
16. Skalický P, Vlasák A, Mládek A, et al. **Role of DESH, callosal angle and cingulate sulcus sign in prediction of gait responsiveness after shunting in iNPH patients**. *J Clin Neurosci* 2021;83:99–107 CrossRef Medline
17. Thavarajasingam SG, El-Khatib M, Vemulapalli K, et al. **Radiological predictors of shunt response in the diagnosis and treatment of idiopathic normal pressure hydrocephalus: a systematic review and meta-analysis.** *Acta Neurochir (Wien)* 2023;165:369–419 CrossRef Medline
18. Reiss-Zimmermann M, Scheel M, Dengl M, et al. **The influence of lumbar spinal drainage on diffusion parameters in patients with suspected normal pressure hydrocephalus using 3T MRI.** *Acta Radiology* 2014;55:622–30 CrossRef Medline
19. van Swieten JC, Koudstaal PJ, Visser MC, et al. **Interobserver agreement for the assessment of handicap in stroke patients.** *Stroke* 1988;19:604–07 CrossRef Medline
20. Feng X, Tustison NJ, Patel SH, et al. **Brain tumor segmentation using an ensemble of 3D U-Nets and overall survival prediction using radiomic features.** *Front Comput Neurosci* 2020;14:25 CrossRef Medline
21. Marmarou A, Bergsneider M, Klinge P, et al. **The value of supplemental prognostic tests for the preoperative assessment of idiopathic normal-pressure hydrocephalus.** *Neurosurgery* 2005;57:S17–28 CrossRef Medline
22. Hallqvist C, Grönstedt H, Arvidsson L. **Gait, falls, cognitive function, and health-related quality of life after shunt-treated idiopathic normal pressure hydrocephalus-a single-center study**. *Acta Neurochir (Wien)* 2022;164:2367–73 CrossRef Medline
23. Tullberg M, Persson J, Petersen J, et al. **Shunt surgery in idiopathic normal pressure hydrocephalus is cost-effective-a cost utility analysis.** *Acta Neurochir (Wien)* 2018;160:509–18 CrossRef Medline
24. Shinoda N, Hirai O, Hori S, et al. **Utility of MRI-based disproportionately enlarged subarachnoid space hydrocephalus scoring for predicting prognosis after surgery for idiopathic normal pressure hydrocephalus: clinical research.** *J Neurosurg* 2017;127:1436–42 CrossRef Medline
25. Kockum K, Lilja-Lund O, Larsson EM, et al. **The idiopathic normal-pressure hydrocephalus Radscale: a radiological scale for structured evaluation.** *Eur J Neurol* 2018;25:569–76 CrossRef Medline
26. Shao M, Han S, Carass A, et al. **Brain ventricle parcellation using a deep neural network: application to patients with ventriculomegaly.** *Neuroimage Clin* 2019;23:101871 CrossRef Medline
27. Rau A, Kim S, Yang S, et al. **SVM-based normal pressure hydrocephalus detection.** *Clin Neuroradiol* 2021;31:1029–35 CrossRef Medline
28. Zhou X, Ye Q, Yang X, et al. **AI-based medical e-diagnosis for fast and automatic ventricular volume measurement in patients with normal pressure hydrocephalus.** *Neural Comput Appl* 2022;35:1–10 CrossRef Medline
29. Tsou CH, Cheng YC, Huang CY, et al. **Using deep learning convolutional neural networks to automatically perform cerebral aqueduct CSF flow analysis.** *J Clin Neurosci* 2021;90:60–67 CrossRef Medline
30. Klinge PM, Brooks DJ, Samii A, et al. **Correlates of local cerebral blood flow (CBF) in normal pressure hydrocephalus patients before and after shunting–a retrospective analysis of [(15)O]H(2)O PET-CBF studies in 65 patients**. *Clin Neurol Neurosurg* 2008;110:369–75
31. Quinn TJ, Dawson J, Walters MR, et al. **Functional outcome measures in contemporary stroke trials.** *Int J Stroke* 2009;4:200–05 CrossRef Medline
32. Grønning R, Jeppsson A, Hellström P, et al. **Association between ventricular CSF biomarkers and outcome after shunt surgery in idiopathic normal pressure hydrocephalus**. *Fluids Barriers CNS* 2023;20:77 CrossRef Medline
33. Levin Z, Leary OP, Mora V, et al. **Cerebrospinal fluid transcripts may predict shunt surgery responses in normal pressure hydrocephalus.** *Brain* 2023;146:3747–59 CrossRef Medline
34. Valsecchi N, Mantovani P, Piserchia VA, et al. **The role of simultaneous medical conditions in idiopathic normal pressure hydrocephalus.** *World Neurosurg* 2022;157:e29–39 CrossRef Medline
35. Koo AB, Elsamadicy AA, Renedo D, et al. **Hospital frailty risk score predicts adverse events and readmission following a ventriculoperitoneal shunt surgery for normal pressure hydrocephalus.** *World Neurosurg* 2023;170:e9–20 CrossRef Medline
36. Davis A, Gulyani S, Manthripragada L, et al. **Evaluation of the effect comorbid Parkinson syndrome on normal pressure hydrocephalus assessment.** *Clin Neurol Neurosurg* 2021;207:106810 CrossRef Medline
37. Giannakopoulos P, Montandon ML, Herrmann FR, et al. **Alzheimer resemblance atrophy index, BrainAGE, and normal pressure hydrocephalus score in the prediction of subtle cognitive decline: added value compared to existing MR imaging markers.** *Eur Radiology* 2022;32:7833–42 CrossRef Medline