



Providing Choice & Value
Generic CT and MRI Contrast Agents



CONTACT REP

AJNR

An Artificial Intelligence Tool for Clinical Decision Support and Protocol Selection for Brain MRI







K.A. Wong, A. Hatef, J.L. Ryu, X.V. Nguyen, M.S. Makary and L.M. Prevedello

This information is current as of July 6, 2025.

AJNR Am J Neuroradiol published online 15 December 2022

<http://www.ajnr.org/content/early/2022/12/15/ajnr.A7736>

An Artificial Intelligence Tool for Clinical Decision Support and Protocol Selection for Brain MRI

 K.A. Wong,  A. Hatef,  J.L. Ryu,  X.V. Nguyen,  M.S. Makary, and  L.M. Prevedello



ABSTRACT

BACKGROUND AND PURPOSE: Protocolling, the process of determining the most appropriate acquisition parameters for an imaging study, is time-consuming and produces variable results depending on the performing physician. The purpose of this study was to assess the potential of an artificial intelligence–based semiautomated tool in reducing the workload and decreasing unwarranted variation in the protocolling process.

MATERIALS AND METHODS: We collected 19,721 MR imaging brain examinations at a large academic medical center. Criterion standard labels were created using physician consensus. A model based on the Long Short-Term Memory network was trained to predict the most appropriate protocol for any imaging request. The model was modified into a clinical decision support tool in which high-confidence predictions, determined by the values the model assigns to each possible choice, produced the best protocol automatically and low confidence predictions provided a shortened list of protocol choices for review.

RESULTS: The model achieved 90.5% accuracy in predicting the criterion standard labels and demonstrated higher agreement than the original protocol assignments, which achieved 85.9% accuracy ($\kappa = 0.84$ versus 0.72, P value $< .001$). As a clinical decision support tool, the model automatically assigned 70% of protocols with 97.3% accuracy and, for the remaining 30% of examinations, achieved 94.7% accuracy when providing the top 2 protocols.

CONCLUSIONS: Our model achieved high accuracy on a standard based on physician consensus. It showed promise as a clinical decision support tool to reduce the workload by automating the protocolling of a sizeable portion of examinations while maintaining high accuracy for the remaining examinations.

ABBREVIATIONS: AI = artificial intelligence; CDS = clinical decision support; SRS = stereotactic radiosurgery; LSTM = Long Short-Term Memory

During protocolling of cross-sectional study requests, radiologists review the study indication and the patient's medical history to determine the examination needed and subsequently set study parameters to best answer the ordering provider's specific clinical question. Protocolling decreases the waste of medical resources, minimizes risks to patients (eg, unnecessary exposure to radiation and/or contrast), and reduces patient inconvenience (eg, callbacks for additional imaging). Time spent on protocolling can vary from a few minutes to several hours a day. Schemmel et al¹ reported that protocolling occupied 6.2% of the workday of a

neuroradiology fellow and contributed to frequent interruptions from image interpretation. While protocolling is an important job of radiologists, it is time-consuming and can benefit from greater automation.

Despite the existence of standardized rules such as the American College of Radiology Appropriateness Criteria,² radiologists often disagree about which protocol is best for a particular study.³ Preferences among radiologists can vary considerably depending on training level and experience and may lead to suboptimal protocol selection. Boland et al³ advocated the standardization of protocolling to improve efficiency and patient safety. Artificial intelligence (AI) can potentially increase both the efficiency and standardization of the process.

The goals of this study were to determine the degree of variability of historical protocol selection relative to a criterion standard based on radiologist consensus and to compare the performance of an AI-based solution against this criterion standard. We also sought to evaluate the automated algorithm as a clinical decision support (CDS) tool by using techniques described previously by

Received March 7, 2022; accepted after revision October 15.

From the Department of Radiology (K.A.W., A.H., J.L.R., X.V.N., M.S.M., L.M.P.), The Ohio State University College of Medicine, Columbus, Ohio; Tri-County Radiologists (A.H.), Newark, Ohio; and ProScan Imaging (J.L.R.), Columbus, Ohio.

Please address correspondence to Luciano Prevedello, MD, MPH, Department of Radiology, The Ohio State University College of Medicine, 395 W 12th Ave, Columbus, OH 43210; e-mail: Luciano.Prevedello@osumc.edu



Indicates article with online supplemental data.

<http://dx.doi.org/10.3174/ajnr.A7736>

Table 1: Label frequencies of the training/validation and test sets^a

	Training/Validation Set	Original Test Set before Protocol Review	Final Test Set after Adjudication
MR imaging brain without/with contrast	11,543 (68.4%)	1274 (67.9%)	1156 (6.2%)
MR imaging brain for MS	2235 (13.2%)	242 (12.9%)	251 (13.4%)
MR imaging tumor	963 (5.7%)	113 (6.0%)	149 (7.9%)
MR imaging brain meningioma follow-up	653 (3.9%)	81 (4.3%)	117 (6.2%)
MR imaging brain dedicated seizure with contrast	519 (3.1%)	65 (3.5%)	65 (3.5%)
MR imaging stealth/Stryker/mask/presurgical planning	436 (2.6%)	43 (2.3%)	45 (2.4%)
MR imaging for gamma knife, brain lab, and SRS	332 (2.0%)	41 (2.2%)	58 (3.1%)
MR imaging cranial nerves III–VI without/with contrast	201 (1.2%)	17 (0.9%)	35 (1.9%)
Total	16,882	1876	1876

^a Data are No. (%).

Kalra et al,⁴ in which cases deemed straightforward were automatically protocolled and more complex studies were sent to radiologists to make the final decision.

MATERIALS AND METHODS

Data Set Collection and Preparation

Brain MR imaging was chosen as the focus of this study for its large number of protocol options. After securing approval with a waiver of informed consent from the institutional review board (The Ohio State University Wexner Medical Center), we obtained a de-identified data set containing all brain MR imaging with and without contrast examinations performed at our institution from January 2015 to January 2017. Each examination included order diagnoses, reason for the examination, order comments, and final assigned protocol. The original data set contained a total of 19,721 examinations. Because our patient population consists almost exclusively of adult patients, no pediatric brain MR imaging protocol was included.

Of the 32 protocols available for this examination in our database, several were specific to research studies or very uncommon clinical scenarios. To focus on common clinical protocols, we narrowed our evaluation to protocols with frequencies of at least 1%, which yielded 9 protocols. A total of 18,758 examinations were included (95.1% of the total). The protocols, in order of frequency, were the following: 1) MR imaging brain without/with contrast; 2) MR imaging brain for MS; 3) MR imaging tumor; 4) MR imaging meningioma follow-up; 5) MR imaging brain–dedicated seizure with contrast; 6) MR imaging stealth/Stryker/mask/presurgical planning; 7) MR imaging brain for gamma knife, brain lab, and stereotactic radiosurgery (SRS); 8) MR imaging cranial nerves III–VI without/with contrast; and 9) MR imaging for brain metastasis. The MR imaging brain without/with contrast and MR imaging for brain metastasis protocols differed only in that perfusion imaging is included in the MR imaging for brain metastasis protocol if the patient reports prior radiation treatment to the technologist. Because the decision to include perfusion is not made at the time of protocoling and MR imaging for brain metastasis was the least common protocol in our data set (1% of all studies), we combined the MR imaging brain without/with contrast and MR imaging for brain metastasis protocols into 1 class. The number of examinations assigned to each protocol is shown in Table 1. Additional details regarding the specific parameters and sequences used in each protocol and a complete list of protocols is shown in the Online Supplemental Data.

The examinations were randomized and divided into a training/validation set containing 16,882 examinations (90%) and a test set containing 1876 examinations (10%). Examinations in the test set were manually protocolled by 2 diagnostic radiology residents (postgraduate year 3 and postgraduate year 2 at the time of review). The residents were blinded to the final protocol assignment but were provided with procedure descriptions (general study ordered, eg, MR imaging brain with and without contrast), coded order diagnosis, reason for examination, order information, order comments, and authorizing provider and department. Criterion standard labels were prepared for the test set by comparing the protocols chosen by the residents. For examinations on which the residents agreed, the resident's protocol choice was taken to be the criterion standard label for the examination. Examinations in which the residents disagreed (207 of 1876 examinations) were reviewed by a board-certified fellowship-trained neuroradiologist, and his protocol choice was taken to be the criterion standard label. Label frequencies of the training/validation set, original test set before protocol review, and final test set after adjudication are available in Table 1.

Word Embeddings and Vocabulary

Word embeddings come from the idea that the meaning of a word can be represented as an array of numbers, or a vector.⁵ Words with similar meanings should correspond to vectors, or embeddings, that lie close together in space. In this work, we used BioWordVec (<https://github.com/ncbi-nlp/BioWordVec>), a set of biomedical-specific word embeddings derived from descriptor terms of medical subject headings and PubMed titles and abstracts.⁶ Its vocabulary contains 2.3 million distinct tokens, of which we used a subset as described later.

Data Preprocessing

For each example in the training/validation set, all associated data, including order diagnoses, reason for the examination, and comments, were concatenated into a single string of text. The text was converted to lowercase. Hyphens were retained, but all other punctuation was removed. Numbers were removed, including ages, dates, and diagnosis codes. Text was split into tokens via whitespace. Tokens from the training/validation set were matched with terms in the BioWordVec vocabulary. Tokens that failed to match (1006 of 7953 unique tokens) were manually reviewed and mapped to a similar term in the BioWordVec vocabulary. The final vocabulary comprised 7102 distinct terms.

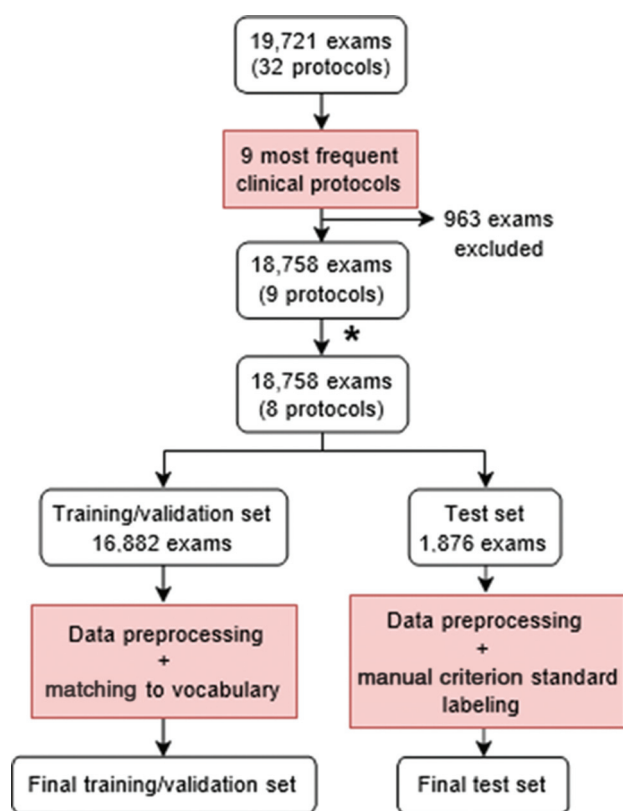


FIG 1. Flow chart describing preparation of training/validation and test sets. The asterisk indicates that due to its infrequency within the data set (1%) and its similarities with the MR imaging brain without/with contrast protocol, the MR imaging for brain metastasis protocol was combined with the MR imaging brain without/with contrast protocol, reducing the total number of protocols from 9 to 8.

Examples in the test set underwent similar preprocessing, but unmatched tokens were not manually reviewed and were instead mapped to the unknown term token. Examinations in the training/validation set contained an average of 25.4 (SD, 19.4) tokens, and those in the test set contained an average of 24.7 (SD, 18.2) tokens. An overview of the data preparation process can be found in Fig 1.

Natural Language-Processing Model

The Long Short-Term Memory (LSTM) network (<https://colah.github.io/posts/2015-08-Understanding-LSTMs/>) was used for the text-classification task. The LSTM is a type of recurrent neural network that uses multiplicative gates to update an internal state, allowing it to retain information during long input sequences.⁷ LSTMs have been shown to perform well on several text-classification benchmarks, particularly when pretrained as a language model and fine-tuned as a text classifier.⁸

To pretrain the LSTM, we assigned a language model around the LSTM to perform an autoregressive sentence-completion task. An embedding layer populated with BioWordVec pretrained word embeddings was placed before the LSTM to match input tokens with their corresponding embeddings. These embeddings were fed into a single bidirectional LSTM layer with a hidden size of 256. Outputs from the LSTM were passed through a fully-connected

layer with output the size of the vocabulary to predict the next token. During training, the training/validation set was randomly split into training (80%) and validation (20%) sets. The language model was trained across 10 epochs of the training set using cross-entropy loss.

Next, the LSTM was fine-tuned on the classification task of predicting the most appropriate imaging protocol of 8 possible choices. Weights from the language model were copied to the classifier model before training, and the last layer was replaced with a fully-connected layer with an output size of 8. Again, the training/validation set was randomly split into training (80%) and validation (20%) sets. Class weights for the training/validation set were computed so that the weight of each class was proportional to the inverse of its prevalence. These class weights were used in the calculation of cross-entropy loss to scale the loss for each class so that loss associated with rarer classes would have greater impact on the weights of the model than loss associated with more prevalent classes. This process allowed the model to compensate for imbalanced class distribution. The classifier model was trained across 5 epochs of the training set and evaluated on the test set with criterion standard labels.

Hyperparameter tuning was performed using a grid search, and optimal values were found to be a batch size of 64, a learning rate of 10^{-7} , an LSTM hidden size of 256, and the number of LSTM layers at 1. See the Online Supplemental Data for an overview of the network architectures. The code used to train and evaluate the model can be found at <https://github.com/kwong22/protocol-brain-mri>.

CDS Tool

Inspired by the work of Kalra et al,⁴ we evaluated the performance of our classifier model as a CDS tool. After receiving input, the model outputs a probability for each possible class. Usually, the class with the greatest probability is chosen as the predicted class. Kalra et al proposed that a threshold be set so that whenever the model produces an output in which the greatest class probability is greater than or equal to the threshold (high confidence), the model enters “automatic” mode and produces the 1 class with the greatest probability. If no probability in the output reaches the threshold (low confidence), the model enters the CDS mode and yields the 3 classes with the greatest probabilities. We applied this idea to our model by testing different thresholds and the number of suggestions for CDS mode.

Tools and Libraries

All code was written in Python (Version 3.8.11; <http://www.python.org>). Word embeddings were loaded using the Gensim library (<https://radimrehurek.com/gensim/>).⁹ Machine learning was performed using PyTorch (<https://pytorch.org/>).¹⁰ The Cohen κ score for interrater reliability was calculated using the scikit-learn library (<https://scikit-learn.org/stable/index.html>).¹¹ The McNemar test of homogeneity was performed using the Statsmodels library (<https://www.statsmodels.org/v0.10.1/>).¹² Model predictions were interpreted using the PyTorch Captum library (<https://github.com/pytorch/captum>)¹³ with the Integrated Gradients method.¹⁴

Table 2: Performance of the classifier model^a

Labels	Weighted Precision	Weighted Recall (Accuracy)	Weighted F1 Score
Criterion standard labels	0.908	0.905	0.905
Original protocol assignments	0.872	0.841	0.850

^aPrecision, recall, and F1 score were calculated for each class, and weighted averages of these metrics were computed using class frequencies as weights. The weighted F1 score represents the average of F1 scores weighted by class frequency.

Actual	WO/W	1061	15	39	10	17	2	0	12
	MS	6	244	0	0	0	0	0	1
	TUMOR	28	0	115	2	1	0	3	0
	MENINGIOMA	4	0	0	113	0	0	0	0
	SEIZURE	1	1	0	0	63	0	0	0
	STEALTH	2	0	1	1	0	40	0	1
	GAMMA	5	0	6	1	0	0	45	1
	CN	15	0	0	3	0	0	0	17
		WO/W	MS	TUMOR	MENINGIOMA	SEIZURE	STEALTH	GAMMA	CN
		Predicted							

FIG 2. Confusion matrix for the LSTM classifier model. Criterion standard labels are on the y-axis. Labels predicted by the model are on the x-axis. Numbers represent the number of examinations in the test set. WO/W indicates MR imaging brain without/with contrast; MS, MR imaging brain for MS; TUMOR, MR imaging tumor; MENINGIOMA, MR imaging brain meningioma follow-up; SEIZURE, MR imaging brain dedicated seizure with contrast; STEALTH, MR imaging stealth/Stryker/mask/pre-surgical planning; GAMMA, MR imaging for gamma knife, brain lab, and SRS; CN, MR imaging cranial nerves III–VI without/with contrast.

Table 3: Contingency table comparing performances of the classifier model and the original protocol assignments on the criterion standard labels^a

	Incorrectly Predicted by Original Assignments	Correctly Predicted by Original Assignments
Incorrectly predicted by model	80	98
Correctly predicted by model	185	1513

^aNumbers represent number of examinations in the test set that were incorrectly/correctly classified by the model and incorrectly/correctly classified by the original protocol assignments. The McNemar test of homogeneity statistic is 26.13; *P* value < .001.

RESULTS

The model achieved weighted F1 scores of 0.905 when predicting the criterion standard labels and 0.850 when predicting the original protocol assignments (Table 2). When predicting the criterion standard labels, the model demonstrated strong performance on 6 of 8 classes, with F1 scores for the top 6 classes ranging from 0.849 to 0.955 (Online Supplemental Data). The 2 classes in which the model performed worst were MR imaging tumor and MR imaging cranial nerves protocols, which had F1 scores of 0.742 and 0.507, respectively. The confusion matrix can be found

in Fig 2. Examples of correct and incorrect model predictions and their associated interpretations via the Integrated Gradients method can be found in the Online Supplemental Data.

The Cohen κ score, a measure of interrater agreement, between the output of the model and the criterion standard labels was 0.842, whereas the interrater agreement between the original protocol assignments and the criterion standard labels was 0.719. In predicting the criterion standard labels, the model (90.5% accuracy) performed significantly better than the original protocol assignments (85.9% accuracy) according to the McNemar test of marginal homogeneity (McNemar χ^2 statistic of 26.13, *P* value < .001) (Table 3).

To evaluate the model as a CDS tool, we varied 2 parameters: the confidence threshold that delineates the automation and CDS modes of the model, and the number of classes (*k*) returned by the CDS mode (Online Supplemental Data). On the lower extreme, with the threshold at 0.5 and *k* at 2, the model demonstrated a weighted recall of 0.925 in the automation mode (applied to 95% of examinations), 0.835 in the CDS mode (applied to the remaining 5% of examinations), and an overall weighted F1 score of 0.921. At the higher extreme, setting the threshold to 0.9 and *k* to 4 resulted in a weighted recall of 0.982 in the automation mode (applied to 48% of examinations), 0.997 in the CDS mode (applied to the remaining 52% of examinations), and an overall weighted F1 score of 0.990.

DISCUSSION

Protocolling decisions made by different radiologists can vary considerably, and previous work has argued for standardization to reduce suboptimal protocol

choices³ associated with unnecessary imaging, increased cost, and patient dissatisfaction. AI, with its ability to analyze large amounts of data and automate tasks in a consistent manner, has the potential to not only increase the accuracy and efficiency of the protocolling process but also to decrease variability. In this study, we compared the performance of the original protocol assignment with that of an AI tool relative to a criterion standard reference based on radiologist consensus. Our AI model showed greater agreement with the criterion standard labels compared with the original protocol assignments, supporting the idea that AI tools

can decrease variability in protocol selection for advanced imaging modalities.

Prior studies have demonstrated the effectiveness of AI in the automation of study protocolling. Brown and Marotta¹⁵ used various machine learning classifiers to select the most appropriate protocol from 13 classes and achieved a recall of 0.83 with a random forest classifier. Kalra et al⁴ evaluated a deep neural network as a CDS tool for protocolling and found that their algorithm protocollered 69% of studies with a weighted recall of 0.951 and, for the remaining 31% of studies, suggested the correct protocol from 3 choices with a weighted recall of 0.915. Our AI model demonstrated an overall strong performance on our criterion standard labels, with high precision (0.908) and recall (0.905).

However, the performance of our model was suboptimal in some categories, particularly MR imaging tumor and cranial nerve protocols. The MR imaging tumor protocol had a low sensitivity of 0.772, indicating a large number of false-negatives, predominantly when the model instead predicted the MR imaging brain without/with contrast protocol. This outcome seems reasonable given that these 2 protocols share the same foundation protocol, differing only in the use of perfusion imaging for MR imaging tumor. The model performed especially poorly on the MR imaging cranial nerves protocol, with a sensitivity of 0.486 and a positive predictive value of 0.531. MR imaging cranial nerves was the second least common protocol in the original data set, suggesting that the poor performance of the model on this class may be due to the lack of data for this class. Further research is needed to determine whether a larger data set would improve performance.

Review of the classification errors of the model revealed that the model learned to associate certain highly predictive terms with protocols, for instance, “gbm” (glioblastoma) for MR imaging tumor; “gamma knife” and “radiation” for MR imaging for gamma knife, brain lab, and SRS; and “facial” for MR imaging cranial nerves. Examinations that mentioned “weakness,” “paresthesia,” and “numbness” were often classified by the model as MR imaging for MS, and “seizure” was often associated with MR imaging dedicated seizure. However, when these highly predictive terms were combined in 1 indication (eg, “possible gbm progression presenting with seizures” or “right facial numbness”), the confidence level of the model decreased, as shown in the Online Supplemental Data. Further research is needed to assess whether a combination of AI and string search methods would benefit performance.

We also evaluated a CDS tool with the goal of reducing the manual workload and increasing protocolling accuracy. The idea behind this tool is that the model will evaluate every study before it reaches any human reviewers. High-confidence predictions result in automatic protocol assignments, while low-confidence predictions present a shortened list of suggested protocols to the appropriate radiologist. The confidence threshold should be low enough that a large portion of examinations are evaluated confidently and automatically by the model, resulting in a noticeable workload reduction for radiologists, and high enough that the model automatically evaluates only examinations about which it is truly confident, ensuring high performance. In addition, *k*, the number of suggested protocols provided by the model in the

CDS mode, should be small enough that the model adequately narrows the range of possibilities and large enough that the optimal protocol choice is almost always included in the suggestions.

Setting our model to protocol all studies in the automation mode results in 90.5% accuracy, but accuracy can be further increased by directing examinations with low-confidence predictions to the CDS mode, in which reviewers are provided a list of *k* protocols from which to choose. Varying *k* from 2 to 4 resulted in accuracies of the CDS mode ranging from 83.5% to 99.7%. A small list of protocols that almost always includes the most appropriate protocol could serve as a definitive resource for reviewers. Increasing the confidence threshold from 0.5 to 0.9 increased the accuracy of automation mode from 92.5% to 98.2% (Online Supplemental Data). However, in doing so, the percentage of studies protocollered automatically fell from 95% to 48%. Thus, this tool can decrease the protocolling workload by a factor of 2 (if 48% of studies are protocollered automatically), with even greater reductions possible depending on the user’s tolerance for automated protocolling errors. This observation illustrates a practical concept in clinical AI implementation in that the AI tool does not need to perform perfectly to improve radiology workflow.

It is difficult to directly compare our tool with that of Kalra et al⁴ because our project scopes differ considerably. While we focused exclusively on brain MR imaging studies, Kalra et al included multiple body regions and imaging modalities. In addition, our model selected from 8 protocol choices, while theirs selected from 108. To adapt our model to make a fair comparison, we selected the minimum *k* for the CDS mode (*k* = 2) and a confidence threshold (threshold = 0.8) to match the 69% rate of high-confidence predictions reported by Kalra et al. With these settings, our model automatically evaluated 70% of examinations with an accuracy of 97.3%; for the remaining 30% of examinations, it included the optimal protocol choice in the top 2 suggestions 94.7% of the time. Compared with the CDS tool created by Kalra et al, our tool achieved higher accuracy in both the automation mode (97.3% versus 95.1%) and the CDS mode (94.7% versus 91.1%).

This study was limited by a relatively low number of cases in certain classes, such as the MR imaging cranial nerves protocol, leading to poor performance by the model. Additional data may increase performance on these classes. Another limitation was the lack of criterion standard labeling of the training data. Because the model was trained on the original assigned protocols, it may have been limited by the variability inherent in those protocol choices. If criterion standard labels were available for the training data, model performance may improve.

In addition, the information provided to the residents and to the attending physician during creation of the criterion standard labels was limited. Key information that would be available through the patient chart, including provider notes, laboratory values, prior imaging examinations, and provider-to-provider communication, was not included in our data set. Future research should investigate how AI performs in a more complex environment that accounts for current patient status and additional information in the electronic medical record.

Furthermore, this study included a limited number of protocol choices. Indeed, only the most frequent protocols were selected,

which simplified the task substantially. Inclusion of more protocols, as will be necessary for real-world applications, will require additional data and model training. Note that imaging study requests and available protocol choices differ among institutions, so clinicians hoping to use our tool will need to retrain it on data from their own institutions for more tailored results.

Finally, the scope of this study was small in that it focused solely on brain MR imaging studies. This focus restricts the ability of the study to generalize to other imaging domains. However, creating a single algorithm to protocol all radiology studies may be impractical. Having separate algorithms for different study types may enable the algorithms to acquire domain-specific knowledge and assign protocols more accurately, just as ours did for brain MR imaging examinations.

CONCLUSIONS

Our AI-based model achieved high accuracy (90.5%) on a reference standard based on physician consensus and has the potential to decrease variability in protocolling by serving as a reliable reference for radiologists during the protocolling process. In addition, our model showed promise as a CDS tool to reduce the workload by automating the protocolling of 70% of examinations with 97.3% accuracy while maintaining 94.7% accuracy for the remaining 30% of examinations when providing the top 2 protocols.

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

REFERENCES

1. Schemmel A, Lee M, Hanley T, et al. **Radiology workflow disruptors: a detailed analysis.** *J Am Coll Radiol* 2016;13:1210–14 [CrossRef Medline](#)
2. American College of Radiology. **ACR Appropriateness Criteria Overview.** <https://www.acr.org/Clinical-Resources/ACR-Appropriateness-Criteria/Overview>. Accessed May 4, 2021
3. Boland GW, Duszak R, Kalra M. **Protocol design and optimization.** *J Am Coll Radiol* 2014;11:440–41 [CrossRef Medline](#)
4. Kalra A, Chakraborty A, Fine B, et al. **Machine learning for automation of radiology protocols for quality and efficiency improvement.** *J Am Coll Radiol* 2020;17:1149–58 [CrossRef Medline](#)
5. Harris ZS. **Distributional structure.** *WORD* 1954;10:146–62 [CrossRef](#)
6. Zhang Y, Chen Q, Yang Z, et al. **BioWordVec, improving biomedical word embeddings with subword information and MeSH.** *Sci Data* 2019;6:52 [CrossRef Medline](#)
7. Hochreiter S, Schmidhuber J. **Long Short-Term Memory.** *Neural Comput* 1997;9:1735–80 [CrossRef Medline](#)
8. Howard J, Ruder S. **Universal language model fine-tuning for text classification.** May 23, 2018. *ArXiv180106146 Cs Stat* <http://arxiv.org/abs/1801.06146>. Accessed April 8, 2021
9. Rehurek R, Sojka P. **Software framework for topic modelling with large corpora.** In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta. May 22, 2010:45–50
10. Paszke A, Gross S, Massa F, et al. **PyTorch: an imperative style, high-performance deep learning library.** In: *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, Canada. December 8–14, 2019
11. Pedregosa F, Varoquaux G, Gramfort A, et al. **Scikit-learn: machine learning in Python.** *J Mach Learn Res* 2011;12:2825–30
12. Seabold S, Perktold J. **Statsmodels: econometric and statistical modeling with Python.** In: *Proceedings of the 9th Python in Science Conference (SciPy 2010)*, Austin, Texas. June 28 to July 3, 2010 [CrossRef](#)
13. Kokhlikyan N, Miglani V, Martin M, et al. **Captum: a unified and generic model interpretability library for PyTorch.** *ArXiv200907896 Cs Stat*. September 16, 2020. <http://arxiv.org/abs/2009.07896>. Accessed February 12, 2022
14. Sundararajan M, Taly A, Yan Q. **Axiomatic attribution for deep networks.** *ArXiv170301365*. June 12, 2017. <http://arxiv.org/abs/1703.01365>. Accessed February 12, 2022
15. Brown AD, Marotta TR. **A natural language processing-based model to automate MRI brain protocol selection and prioritization.** *Acad Radiol* 2017;24:160–66 [CrossRef Medline](#)