Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Machine Learning and Deep Learning Models for Automated Protocoling of Emergency Brain MRI Using Text from Clinical Referrals

Heidi J. Huhtanen, MD[1] • *Mikko J. Nyman, MD, PhD*[1] • *Antti Karlsson, PhD*[2] • *Jussi Hirvonen, MD, PhD*[1]

Author affiliations, funding, and conflicts of interest are listed at the end of this article. See also commentary by Strotzer in this issue.

Radiology: Artificial Intelligence 2025; 7(3):e230620 • https://doi.org/10.1148/ryai.230620 • Content codes: AI NR MR

Purpose: To develop and evaluate machine learning and deep learning–based models for automated protocoling of emergency brain MRI scans based on clinical referral text.

Materials and Methods: In this single-institution, retrospective study of 1953 emergency brain MRI referrals from January 2016 to January 2019, two neuroradiologists labeled the imaging protocol and use of contrast agent as the reference standard. Three machine learning algorithms (naive Bayes, support vector machine, and XGBoost) and two pretrained deep learning models (Finnish bidirectional encoder representations from transformers [BERT] and generative pretrained transformer [GPT]–3.5 [GPT-3.5 Turbo; Open AI]) were developed to predict the MRI protocol and need for a contrast agent. Each model was trained with three datasets (100% of training data, 50% of training data, and 50% plus augmented training data). Prediction accuracy was assessed with a test set.

Results: The GPT-3.5 models trained with 100% of the training data performed best in both tasks, achieving an accuracy of 84% (95% CI: 80, 88) for the correct protocol and 91% (95% CI: 88, 94) for the contrast agent. BERT had an accuracy of 78% (95% CI: 74, 82) for the protocol and 89% (95% CI: 86, 92) for the contrast agent. The best machine learning model in the protocol task was XGBoost (accuracy, 78%; 95% CI: 73, 82), and the best machine learning models in the contrast agent task were support vector machine and XGBoost (accuracy, 88%; 95% CI: 84, 91 for both). The accuracies of two nonneuroradiologists were 80%–83% in the protocol task and 89%–91% in the contrast medium task.

Conclusion: Machine learning and deep learning models demonstrated high performance in automatic protocoling of emergency brain MRI scans based on text from clinical referrals.

Supplemental material is available for this article.

Published under a CC BY 4.0 license.

Protocoling of incoming imaging studies is estimated to take 3.5%–6.2% of radiologists' time (1,2). In emergency radiology, protocoling is a frequent cause of interruptions, which may increase cognitive workload and risk of errors (3,4). However, careful protocoling is essential for answering the clinical question. Errors are common in protocoling (5,6), which is often done by less experienced residents or nonsubspecialist radiologists in an emergency setting. During the last decade, advancements in artificial intelligence and natural language processing (NLP) (7) have opened new possibilities for streamlining the protocoling process.

Although most NLP research in radiology has focused on analyzing radiology reports and communicating critical findings (8,9), there is growing interest in applying NLP for automatic protocol selection (10–21). Promising results have been shown with traditional machine learning (ML) algorithms such as support vector machine (SVM), random forest, and XGBoost (10–13). Deep learning (DL)–based large pretrained language models, such as bidirectional encoder representations from transformers (BERT) (22), have also demonstrated good performance in protocoling (14,15). Specific versions of BERT tailored for radiology might further enhance the results (15,23), although these have been limited to English language. Larger language models, such as the generative pretrained transformer (GPT) (24) from OpenAI, have recently gained attention for their ability to learn complex tasks with limited data and achieve humanlike performance in tasks such as text generation, emotional awareness, or medical examinations (24–26). However, to our knowledge, there is currently only one published study investigating GPT's potential in automatic protocoling (16).

In this study, we evaluate both ML and DL models in automatic protocoling of emergency brain MRI scans using text from clinical referrals. We tested three ML models (naive Bayes, SVM, and XGBoost) and two DL models (BERT and GPT-3.5). Because curating large specialist-annotated medical datasets is resource intensive, we also assessed the performance of the models when trained with datasets of varying sizes.

Materials and Methods

We obtained permission for this study from the Hospital District of Southwest Finland. Review by the institutional review board or ethics committee was not required because registry-based retrospective studies of existing data in Finland are exempted from ethical approval by law and are only subject to hospital district permission. A waiver for written patient consent was not sought for the same reason. The study was conducted in accordance with the Declaration of Helsinki.

Data

Protocoling of emergency brain MRI scans using referral texts was chosen as the study target because MRI is used daily at our emergency department, and the annual number of emergency brain MRI referrals has almost doubled from 1182 in 2014 to 2089 in 2018 (Fig 1). In our institution, "emergency

This copy is for personal use only. To order printed copies, contact *reprints@rsna.org*

Abbreviations

BERT = bidirectional encoder representations from transformers, DL = deep learning, GPT = generative pretrained transformer, ML = machine learning, NLP = natural language processing, SVM = support vector machine

Summary

Both traditional machine learning models and newer deep learning models, like bidirectional encoder representations from transformers and generative pretrained transformer–3.5 (Open AI; GPT-3.5 Turbo), performed well in automatic protocoling of emergency brain MRI scans based on text from clinical referrals.

Key Points

- Machine learning (support vector machine, XGBoost, and naive Bayes) and deep learning (bidirectional encoder representations from transformers [BERT] and generative pretrained transformer [GPT]–3.5 [Open AI; GPT-3.5 Turbo]) models were developed to predict the emergency brain MRI protocol and need for a contrast agent based on text from clinical referrals.
- The GPT-3.5 models trained with the large nonaugmented dataset achieved the best results (accuracy, 84% for predicting the protocol and 91% for determining the need for a contrast agent), with performance comparable to that of nonneuroradiologists (accuracy, 80%–83% and 89%–91%, respectively).
- The BERT models and the best machine learning models also demonstrated high performance, with accuracies of 78% in predicting the MRI protocol and 89% and 88%, respectively, in determining the need for a contrast agent.

Keywords

Natural Language Processing, Automatic Protocoling, Deep Learning, Machine Learning, Emergency Brain MRI

referrals" refers to requests for urgent imaging within 1–3 days. We retrospectively collected a random sample of 2000 Finnish emergency brain MRI referrals from January 2016 to January 2019 from Turku University Hospital. Forty-seven referrals were excluded for various reasons: insufficient information (n = 2), nonemergency MRI or preoperative planning (n = 9), scientific projects (n = 32), or wrong study code (n = 4). The data were anonymized by removing identifiable attributes, including patient names, personal identification numbers, and examination dates. The referral texts contained free text written in a single field, and their average word count was approximately 50 words.

The included referrals (n = 1953) were classified into suitable imaging protocols and use or no use of a gadolinium-based contrast agent by two fellowship-trained neuroradiologists (M.J.N., 16 years of experience, and J.H., 12 years of experience) in consensus, regarded as a better reference standard than original instructions from radiologists of varying experience. There were 12 different protocol classes and two contrast agent classes (contrast and noncontrast) (Fig 2). Detailed information about the protocol classes used can be found in Table S1.

The data were split into training (80%) and test sets (20%) by stratified sampling to maintain consistent protocol class proportions. Inspecting the datasets showed the contrast agent class proportions had also stayed roughly equal. To study the effect of training set size for model performance, we randomly sampled a 50% smaller dataset (called "small") from the original "large" training set.

Preprocessing

We preprocessed the referral texts by simplifying white spaces and removing unnecessary automatic sentences generated by the radiology information system. The ML models required texts to be further preprocessed by removing punctuation and "stop-words" and changing conjugated words into their base form. We expanded the small dataset with data augmentation to create a larger "augmented" dataset. Preprocessing and augmentation were done with Python (version 3.7.11; Python Software Foundation), and details can be found in Appendix S1.

The training sets were split into five stratified folds by protocol class, with original referrals and their augmentations kept in the same fold. We checked that each fold's contrast agent class distribution remained close to the training set's baseline. During training, augmented referrals were excluded from the validation fold to ensure validation relied solely on original referrals.

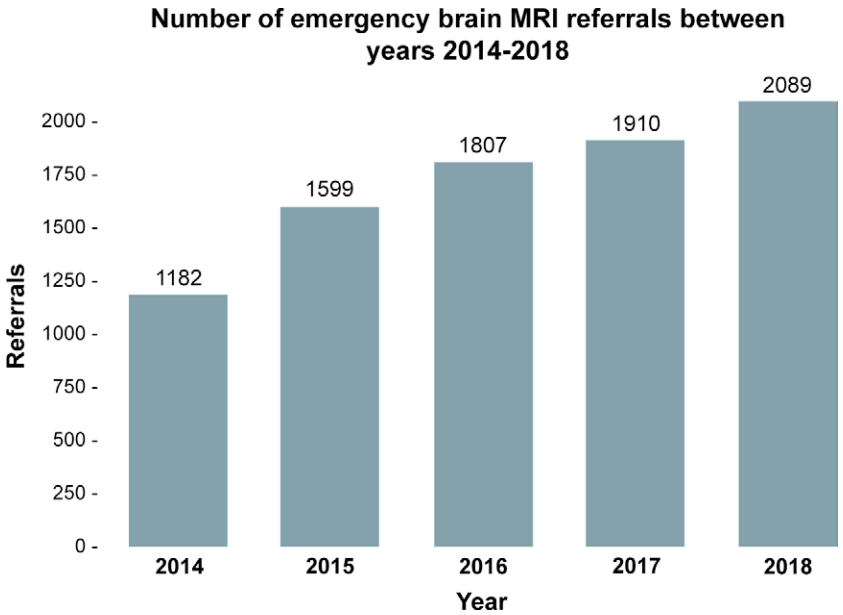

Figure 1: Bar graph of the number of emergency brain MRI referrals in the institution from January 2014 to December 2018. The emergency brain MRI referrals include both patients from the emergency department and patients from hospital wards that need imaging urgently.

Model Architectures

For the baseline, we chose three types of ML models: an SVM and an XGBoost algorithm, both of which have shown promise in earlier studies (10,11,13), and a naive Bayes model, a common ML model not previously studied in automatic protocoling. As a vectorizer, we used an n-gram model that turns text into numerical vectors based on combinations of n consecutive letters.

For DL models, we tested BERT (22), specifically a pretrained Finnish version called FinBERT (27), and GPT (24), specifically the OpenAI GPT-3.5 Turbo language model (30), which was available for fine-tuning. We did not have access to GPT-4.

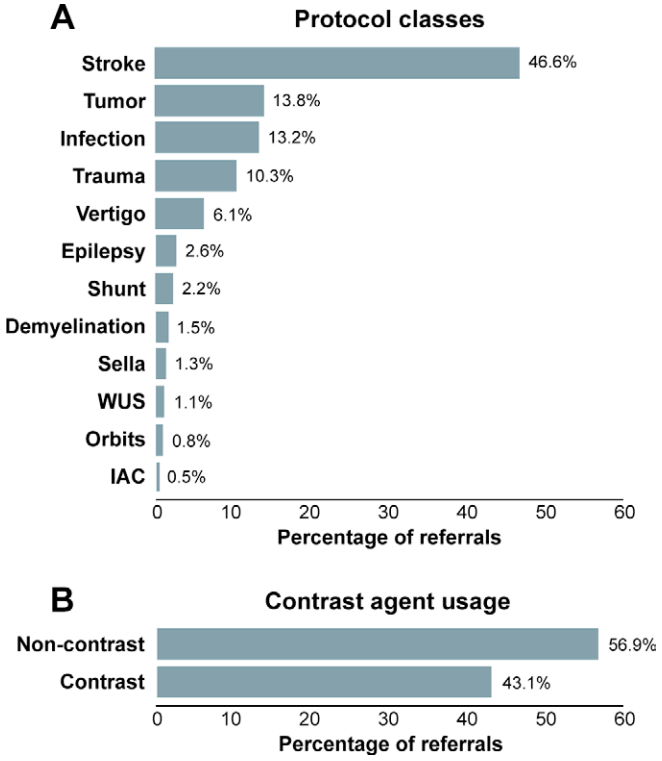

Figure 2: Bar graphs of overall class distributions for the whole dataset (both training and test data). **(A)** Distribution of protocol classes. **(B)** Distribution of contrast agent use. IAC = internal auditory canal, WUS = wake-up stroke.

Model Training

Models were trained separately to predict the MRI protocol and the need for a contrast agent. The ML and BERT models were trained using Python on a local computer equipped with a graphics processing unit (Nvidia Tesla V100). The source code is available at GitHub *(https://github.com/turku-rad-ai/finnish-brain-mri-protocoling)*.

Optimal hyperparameters for ML models were determined via fivefold cross-validation, and final models were trained using these. For BERT models, extensive hyperparameter tuning was limited by computational resources, and hyperparameters were chosen based on preliminary rounds of fivefold cross-validation. Final models were trained for seven epochs using the first data fold (20%) for validation and the rest (80%) for training, selecting the model from the epoch with the lowest validation loss. Details on the chosen hyperparameters for different models are in Appendix S1.

Unlike the other models, fine-tuning and testing the GPT-3.5 models required using OpenAI's web interface and application programming interface on OpenAI's platform. The train-validation split was also done using the platform's tool. GPT-3.5 Turbo (version 0125) was used as a base model and trained using recommended default hyperparameters. Additional information about the data privacy aspects and the prompt engineering can be found in Appendix S1.

Model Evaluation and Statistical Analysis

Model evaluation and statistical analyses were performed using Python and R software (version 4.3.1; R Foundation). Performance was evaluated on the test set using accuracy, F1 score, precision, and recall. The 95% CIs were computed with bootstrapping (10 000 samples) using the boot R library (version 1.3.28.1). For the protocol models, macro (classes have equal value) and weighted (classes are weighted by size) averages of F1 score, precision, and recall were calculated due to class imbalances. For the contrast agent models, binary metrics were calculated defining the contrast class as positive. Accuracies were compared using the McNemar test with Bonferroni correction (in both tasks, the number of pairwise comparisons between the models was 20; the level of significance after Bonferroni correction was $P < .0025$). We did not analyze area under the receiver operating characteristic curve because it is preferably calculated using predicted class probability estimates rather than just predicted labels, and scores from probabilities are not directly comparable to those from labels; also, GPT does not produce these outputs. We qualitatively reviewed examples of correct and incorrect predictions made by the models.

Nonneuroradiologist Performance Analysis

For comparison, two emergency radiologists (radiologist 1, 14 years of experience, and radiologist 2, 15 years of experience), who perform protocoling of emergency MRI referrals daily in their work, independently reviewed the test set. Their performances were analyzed and accuracies compared with the large models using the McNemar test with Bonferroni correction (the number of comparisons was 10 for each task, the protocol and the contrast agent task; the level of significance after Bonferroni correction was $P < .005$).

Results

Dataset Characteristics

Class frequencies and distribution for the training and test sets are listed in Table 1. The large training set had 1563 referrals, the small dataset 783 referrals, and the augmented dataset 2674 referrals. The test set had 390 referrals. The stroke protocol class was most common (46.6% for large, small, and test sets), followed by tumor (13.8%) and infection (13%). The augmented dataset's class distribution varied slightly due to attempts to lessen the class imbalances, but the stroke class remained the largest. The contrast medium usage distribution differed minimally across datasets, the positive class varying between 42.7% and 47.4% and the negative class varying between 52.6% and 57.3%.

Protocol Models

The protocol models' results are shown in Table 2. Overall, the models performed similarly, and pairwise comparisons between the models yielded mostly statistically nonsignificant results (Fig S1A).

Among models trained on the large dataset, GPT-3.5 achieved the highest mean results with an accuracy of 84% (95% CI: 80, 88), macro F1 score of 0.77 (95% CI: 0.61, 0.84), and weighted F1 score of 0.84 (95% CI: 0.80, 0.87). The only statistically significant differences between the accuracies of the large models were between GPT-3.5 and SVM (84% vs 76%; $P = .001$; Bonferroni-corrected $P = .0025$) and between GPT-3.5 and naive Bayes (84% vs 75%; $P < .001$; Bonferroni-corrected $P = .0025$).

Table 1: Training and Test Dataset Characteristics

Variable	Large Training Dataset (*n* = 1563)	Small Training Dataset (*n* = 783)	Small plus Augmented Dataset (*n* = 2674)	Test Dataset (*n* = 390)
Average referral word count	51	52	51	53
Protocols				
Demyelination	24 (1.5)	10 (1.3)	60 (2.2)	6 (1.5)
Epilepsy	40 (2.6)	16 (2.0)	96 (3.6)	10 (2.6)
Internal auditory canal	7 (0.4)	6 (0.8)	36 (1.3)	2 (0.5)
Infection	206 (13.2)	104 (13.3)	416 (15.6)	52 (13.3)
Orbits	13 (0.8)	9 (1.1)	54 (2.0)	3 (0.8)
Sella	20 (1.3)	13 (1.7)	78 (2.9)	5 (1.3)
Shunt	35 (2.2)	13 (1.7)	78 (2.9)	8 (2.1)
Stroke	728 (46.6)	370 (47.3)	740 (27.7)	182 (46.7)
Trauma	161 (10.3)	83 (10.6)	415 (15.5)	40 (10.3)
Tumor	216 (13.8)	102 (13.0)	408 (15.3)	54 (13.8)
Vertigo	96 (6.1)	49 (6.3)	245 (9.2)	24 (6.2)
Wake-up stroke	17 (1.1)	8 (1.0)	48 (1.8)	4 (1.0)
Contrast medium used				
Yes	668 (42.7)	355 (45.3)	1268 (47.4)	173 (44.4)
No	895 (57.3)	428 (54.7)	1406 (52.6)	217 (55.6)

Note.—Data are presented as numbers with percentages in parentheses.

Table 2: Performance Results for the Protocol Models

Model	Accuracy (%)	Macro F1	Weighted F1	Precision*	Recall*
Naive Bayes					
Large	75 (70, 79)	0.49 (0.37, 0.57)	0.73 (0.67, 0.77)	0.74 (0.68, 0.78)	0.75 (0.70, 0.79)
Small	74 (70, 79)	0.48 (0.36, 0.55)	0.71 (0.66, 0.76)	0.74 (0.67, 0.78)	0.74 (0.70, 0.78)
Augmented	77 (73, 82)	0.55 (0.43, 0.62)	0.75 (0.71, 0.80)	0.77 (0.72, 0.81)	0.77 (0.73, 0.82)
SVM					
Large	76 (72, 81)	0.52 (0.41, 0.60)	0.75 (0.70, 0.79)	0.76 (0.70, 0.80)	0.76 (0.72, 0.81)
Small	77 (72, 81)	0.53 (0.42, 0.59)	0.75 (0.70, 0.79)	0.75 (0.71, 0.80)	0.77 (0.72, 0.81)
Augmented	76 (72, 81)	0.52 (0.42, 0.59)	0.75 (0.71, 0.80)	0.76 (0.71, 0.80)	0.76 (0.72, 0.81)
XGBoost					
Large	78 (73, 82)	0.63 (0.50, 0.70)	0.76 (0.72, 0.81)	0.78 (0.74, 0.82)	0.78 (0.73, 0.82)
Small	76 (71, 80)	0.51 (0.38, 0.59)	0.73 (0.68, 0.78)	0.78 (0.71, 0.82)	0.76 (0.71, 0.80)
Augmented	74 (70, 78)	0.58 (0.42, 0.66)	0.73 (0.68, 0.77)	0.75 (0.70, 0.80)	0.74 (0.70, 0.78)
BERT					
Large	78 (74, 82)	0.35 (0.31, 0.39)	0.75 (0.70, 0.80)	0.73 (0.67, 0.78)	0.78 (0.74, 0.82)
Small	69 (65, 74)	0.25 (0.23, 0.26)	0.63 (0.58, 0.69)	0.59 (0.53, 0.65)	0.69 (0.65, 0.74)
Augmented	77 (73, 81)	0.54 (0.43, 0.61)	0.76 (0.71, 0.80)	0.76 (0.72, 0.81)	0.77 (0.73, 0.81)
GPT-3.5					
Large	84 (80, 88)	0.77 (0.61, 0.84)	0.84 (0.80, 0.87)	0.84 (0.80, 0.88)	0.84 (0.80, 0.88)
Small	82 (78, 86)	0.66 (0.54, 0.73)	0.81 (0.77, 0.85)	0.82 (0.77, 0.86)	0.82 (0.78, 0.86)
Augmented	84 (80, 88)	0.74 (0.61, 0.81)	0.85 (0.81, 0.88)	0.86 (0.83, 0.90)	0.84 (0.80, 0.88)
Radiologist					
Radiologist 1	80 (76, 84)	0.71 (0.61, 0.77)	0.81 (0.76, 0.84)	0.82 (0.79, 0.86)	0.81 (0.76, 0.84)
Radiologist 2	83 (79, 87)	0.76 (0.62, 0.82)	0.83 (0.79, 0.87)	0.84 (0.81, 0.88)	0.83 (0.79, 0.87)

Note.—Data in parentheses are 95% CIs. BERT = bidirectional encoder representations from transformers, GPT = generative pretrained transformer, SVM = support vector machine.
* Precision and recall are weighted averages.

BERT had an accuracy of 78% (95% CI: 74, 82), macro F1 score of 0.35 (95% CI: 0.31, 0.39), and weighted F1 score of 0.75 (95% CI: 0.70, 0.80). Among ML models, XGBoost performed the best: 78% (95% CI: 73, 82), 0.63 (95% CI: 0.50, 0.70), and 0.76 (95% CI: 0.72, 0.81) for accuracy, macro F1 score, and weighted F1 score, respectively.

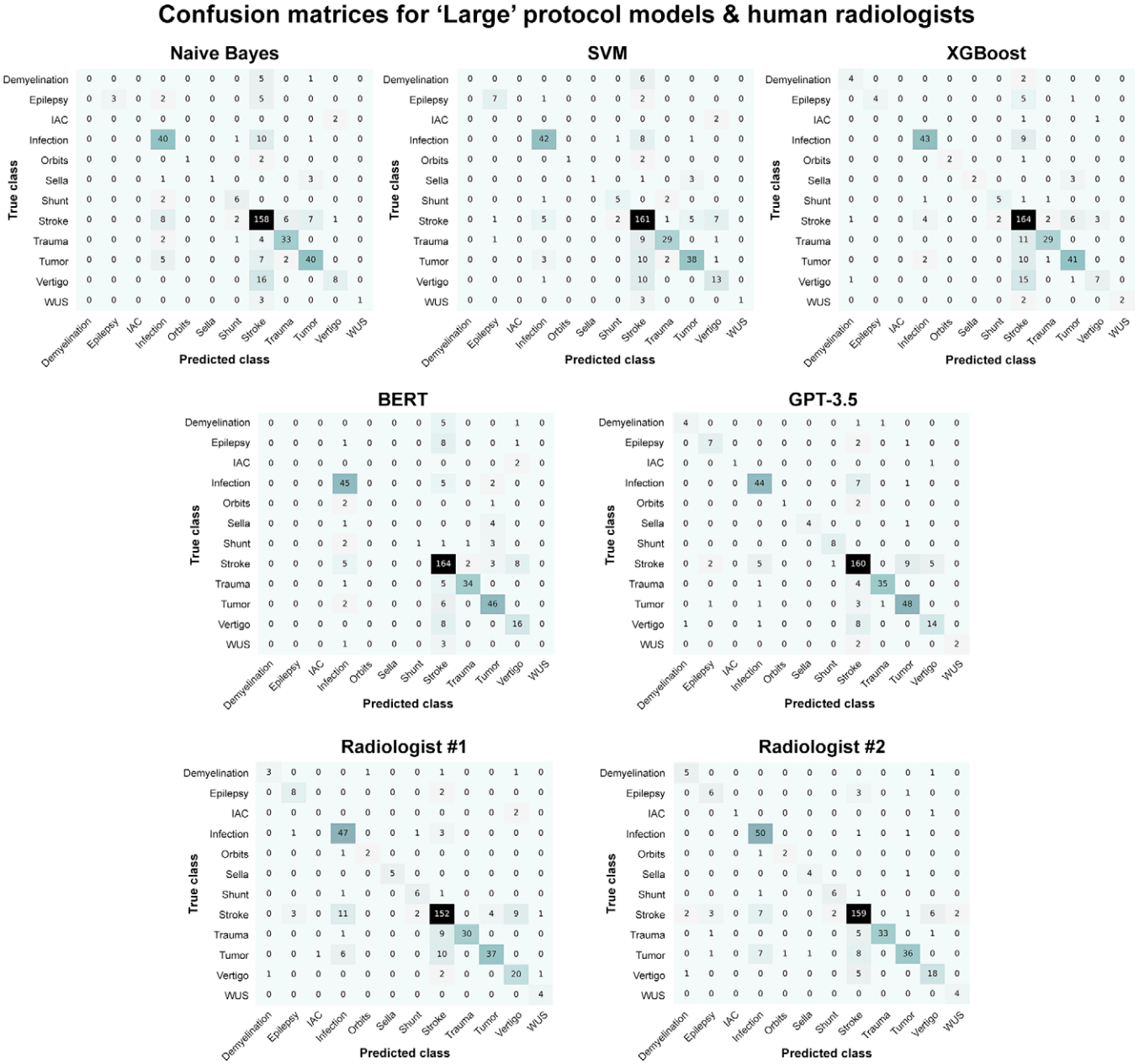

Figure 3: Confusion matrices for the protocol models trained with the large dataset and the radiologists. BERT = bidirectional encoder representations from transformers, GPT = generative pretrained transformer, IAC = internal auditory canal, SVM = support vector machine, WUS = wake-up stroke.

When comparing performance of models trained with large versus small datasets, only BERT showed a statistically significant drop in accuracy (78% vs 69%; $P < .001$; Bonferroni-corrected $P = .0025$). Between the models trained on large versus augmented datasets, there was no evidence of difference ($P > .0025$). Among both the small and augmented models, GPT-3.5 had highest accuracies of 82% (95% CI: 78, 86) and 84% (95% CI: 80, 88), respectively.

All models had higher weighted F1 scores than macro F1 scores. This result implies that the models predicted larger classes better than smaller ones, evident also from the confusion matrices (Figs 3, S2) and the class-specific performance metrics (Table S2).

We manually reviewed individual referrals that were most often associated with correct or incorrect protocol predictions by the models (examples shown in Table S3). Although referrals associated with incorrect model predictions were heterogeneous, many described nonspecific symptoms or otherwise clinically unclear situations. Yet referrals associated with correctly predicted protocols were more straightforward, with clear clinical questions and fewer confounding variables.

Contrast Agent Models

As in the protocol task, the contrast agent models performed evenly (Fig 4), with no evidence of differences in accuracy between large models (Fig S1B). The large GPT-3.5 had the highest mean accuracy of 91% (95% CI: 88, 94) and F1 score of 0.89 (95% CI: 0.86, 0.93). BERT had an accuracy of 89% (95% CI: 86, 92) and F1 score of 0.87 (95% CI: 0.84, 0.91). Among ML models, accuracies and F1 scores were 86% (95% CI: 82, 89) and 0.83 (95% CI: 0.79, 0.88) for naive Bayes,

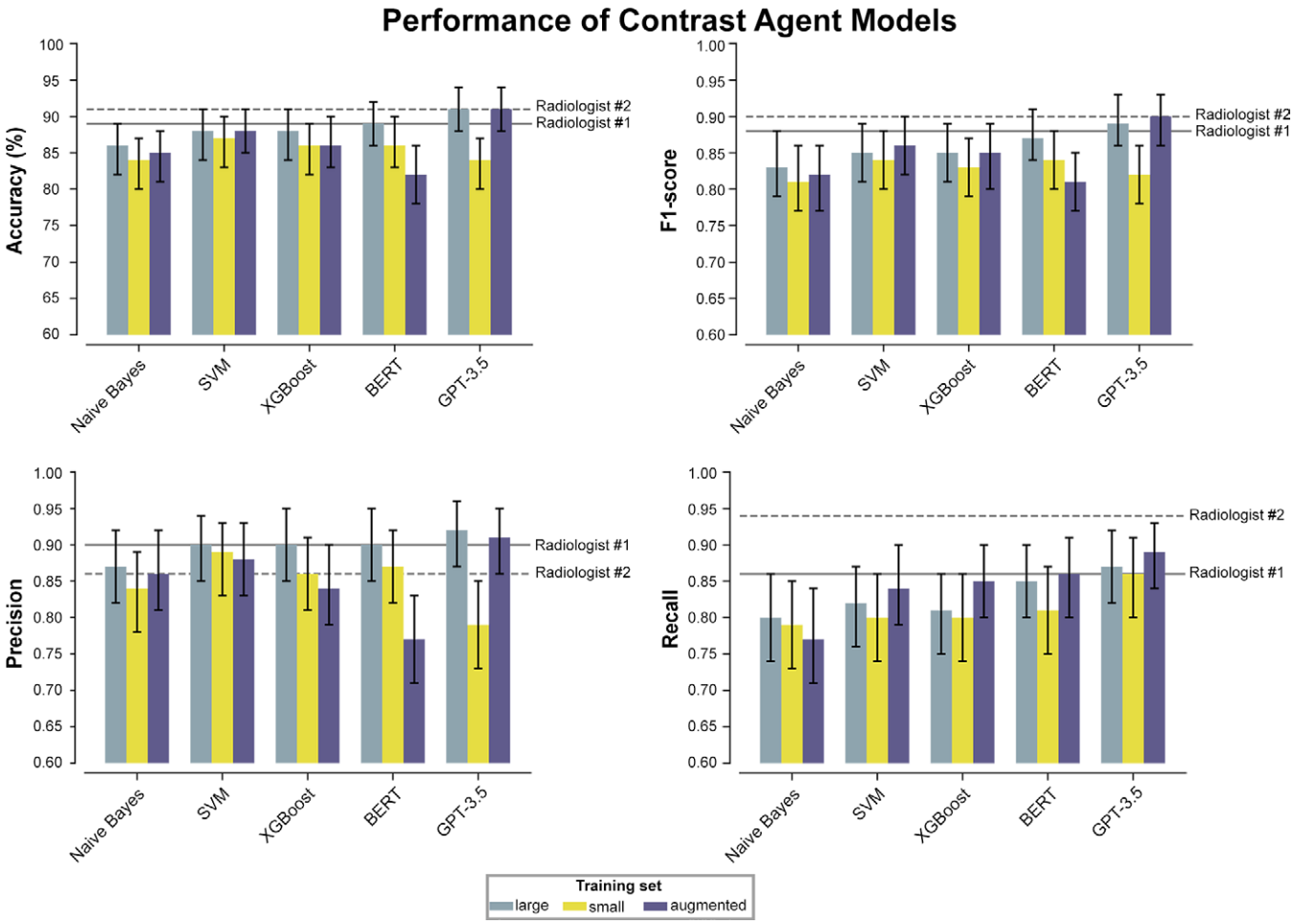

Figure 4: Bar graphs show performance of contrast agent models trained with different datasets. Error bars represent 95% CIs. BERT = bidirectional encoder representations from transformers, GPT = generative pretrained transformer, SVM = support vector machine.

88% (95% CI: 84, 91) and 0.85 (95% CI: 0.81, 0.89) for SVM, and 88% (95% CI: 84, 91) and 0.85 (95% CI: 0.81, 0.89) for XGBoost, respectively. Confusion matrices are shown in Figures 5 and S3.

There was no evidence of differences in accuracy between models trained with large versus small datasets, except for GPT-3.5 (91% vs 84%; $P < .001$; Bonferroni-corrected $P = .0025$). Interestingly, among the small models, GPT-3.5 and naive Bayes had the lowest accuracies (0.84; [95% CI: 0.80, 0.87] for both models). SVM had the highest accuracy of 0.87 (95% CI: 0.83, 0.90). When using data augmentation, the models performed similarly compared with the large models, except for the augmented versus large BERT (0.82 vs 0.89; $P < .001$; Bonferroni-corrected $P = .0025$). Detailed results are shown in Figure S1B.

Examples of correctly and incorrectly predicted protocols based on clinical referrals are shown in Table S4.

Nonneuroradiologist Performance

The results for emergency radiologists are shown in Table 2 and Figure 4, and confusion matrices for their predictions can be found in Figures 3 and 5. Both radiologists performed similarly. In the protocol task, radiologist 1 had an accuracy of 80% (95% CI: 76, 84), macro F1 score of 0.71 (95% CI: 0.61, 0.77), and weighted F1 score of 0.81 (95% CI: 0.76, 0.84). Radiologist 2 had an accuracy of 83% (95% CI: 79, 87), macro F1 score of 0.76 (95% CI: 0.62, 0.82), and weighted F1 score of 0.83 (95% CI: 0.79, 0.87). In the contrast agent task, the accuracies and F1 scores were 89% (95% CI: 86, 92) and 0.88 (95% CI: 0.84, 0.91) for radiologist 1 and 91% (95% CI: 87, 93) and 0.90 (95% CI: 0.86, 0.93) for radiologist 2, respectively. When comparing radiologists to the large models, the only statistically significant difference was between the naive Bayes protocol model and radiologist 2 (accuracy, 75% vs 83%; $P < .001$; Bonferroni-corrected $P = .005$). For the rest, there was no evidence of differences (Fig S4). Exact times were not measured, but radiologists estimated that it took them 5 hours (radiologist 1) and 6 hours (radiologist 2) to review the test set, while artificial intelligence models analyzed the test set in seconds to minutes (exact times listed in Appendix S1).

Discussion

We evaluated NLP models for assigning the correct imaging protocol and determining the need for a contrast agent based on text from emergency brain MRI referrals. We showed that although both ML and DL models demonstrated promising results, GPT-3.5 outperformed them, achieving an accuracy of 84% for protocol selection and 91% for predicting the need for a contrast agent. Moreover, the performance of our best models was on the same level as that of experienced emergency radiologists.

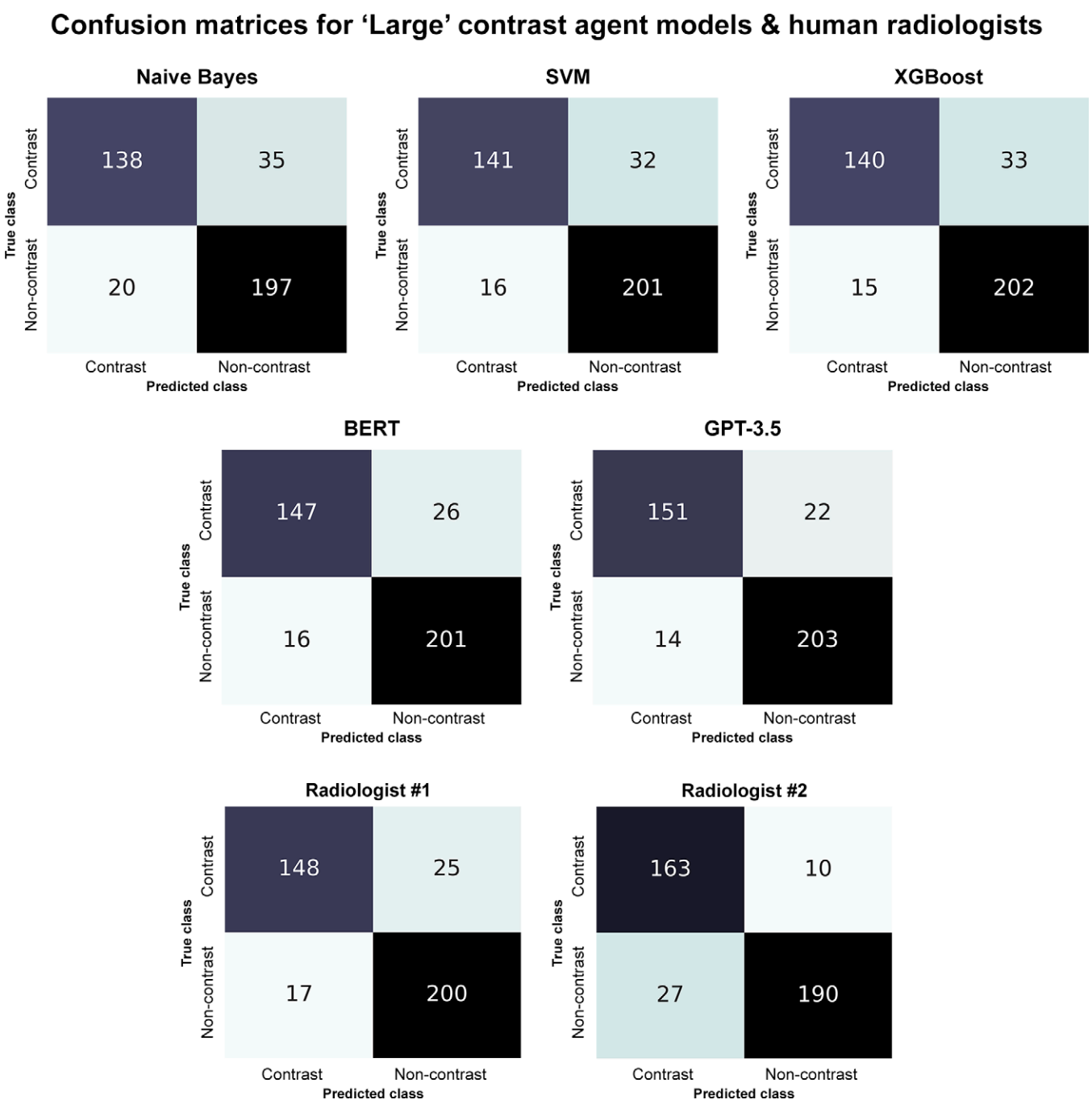

Figure 5: Confusion matrices for the contrast agent models trained with the large dataset and the radiologists. BERT = bidirectional encoder representations from transformers, GPT = generative pretrained transformer, SVM = support vector machine.

DL models (BERT and GPT-3.5) trained with the large dataset had slightly better results than those trained with the small dataset. GPT might still benefit from more training data despite its design as a few-shot learner (24), possibly due to the specialized vocabulary in Finnish MRI referrals, differing considerably from its pretraining data. For most DL models, using data augmentation to upsample the small dataset yielded similar results to the large models, supporting using augmentation when data are scarce. Meanwhile, ML models showed variable results across different training datasets.

Overall, the models performed better at predicting the need for a contrast agent than choosing the protocols, which was expected because the multiclass protocol task was more complex than the binary contrast agent task. Challenges in the protocol task included classifying small classes and cases with nonspecific symptoms like vertigo. Previous studies have shown that MRI outcomes for patients with vertigo or nontraumatic headache vary (28,29), complicating the assignment of a definitive protocol. These insights underscore the complexity of protocoling in diverse clinical situations, in which the correct protocol cannot always be determined solely from the referral.

Although GPT-3.5 performed best, data privacy reasons might limit its clinical usability because the model is used via an application programming interface and cannot be run in house. Thus, BERT or ML models could be preferred.

Our results align with prior research in automated protocoling, though differing study settings limit comparisons. One study (16) tested GPT-4 in determining modality, body part, and contrast agent usage and found promising agreement with the radiologist (85%–100%). BERT achieved an accuracy of 83% in protocoling musculoskeletal MRI referrals (14) and a weighted F1 score of 0.84 in protocoling body CT referrals (15). Protocoling brain MRI referrals has been 83% accurate with random forest (10), 85% with XGBoost (11), and 90.5% with a DL-based long short-term memory network model (17). Our smaller training sample (1563 compared with 16 882 for Wong et al [17]) and use of non-English language might explain some variance. Other studies on various targets showed accuracies of 69%–95% (12,13,17–21), consistent with our results.

Our study had limitations. Our data were retrospective and had an imbalance between the protocol classes. The scarcity of data did not allow assessing model performance on less common classes. The stratified sampling of the training and testing sets partly explains the satisfactory model performance, and performance on data with different class distributions remains untested. We tested the models with only 100% and 50% original datasets, limiting our conclusions on how the dataset size affects performance. Using data from only one hospital may not be a major limitation, as different institutions have different practices for protocoling, and models should also be customized to each hospital's needs. Similarly, using only Finnish data, our approach and results are similar to those using other languages, confirming robustness. We did not analyze the clinical impact of our models' predictions—for example, would the predicted protocol have sufficed in answering the clinical question. Additionally, the emergency radiologists reviewing the test set did not score perfectly, suggesting that multiple protocols may be justified for some cases. Artificial intelligence models may not completely automate the protocoling process yet, considering the complexity of clinical cases, insufficient or incorrect referral information, and additional verbal information radiologists might receive. Still, even partial artificial intelligence automation could reduce interruptions for radiologists.

In conclusion, we found potential for both ML- and DL-based NLP algorithms in automated protocoling of emergency brain MRI scans based on clinical referral text. GPT-3.5 showed the highest performance, but BERT and traditional ML models remain pertinent choices when an in-house solution is preferred. These models could streamline the radiologists' workflow and assist less experienced radiologists. Further research is needed for prospective validation and clinical implementation.

Author affiliations:

[1] Department of Radiology, Turku University Hospital & University of Turku, Kiinamyllynkatu 4-8, 20521 Turku, Finland

[2] Department of Radiology, University of Turku, Turku, Finland, and Pihlajalinna Turku, Turku, Finland

Received December 21, 2023; revision requested February 21, 2024; revision received January 11, 2025; accepted February 4.

Address correspondence to: H.J.H. (email: hejohuh@utu.fi).

Funding: H.J.H. received funding for this study from the Emil Aaltonen Research Foundation (Emil Aaltosen Säätiö, grant no. 230049) and from the Radiological Society of Finland.

Acknowledgment: The authors used OpenAI's ChatGPT (version 4) to assist with grammar and language when writing this article.

Author contributions: Guarantors of integrity of entire study, **H.J.H., M.J.N., J.H.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, all authors; experimental studies, **H.J.H., A.K.**; statistical analysis, **H.J.H., M.J.N., A.K.**; and manuscript editing, all authors

Disclosures of conflicts of interest: **H.J.H.** No relevant relationships. **M.J.N.** No relevant relationships. **A.K.** No relevant relationships. **J.H.** Research grant to institution, Sigrid Jusélius Foundation.

References

1. Dhanoa D, Dhesi TS, Burton KR, Nicolaou S, Liang T. The evolving role of the radiologist: the Vancouver workload utilization evaluation study. J Am Coll Radiol 2013;10(10):764–769.
2. Schemmel A, Lee M, Hanley T, et al. Radiology Workflow Disruptors: A Detailed Analysis. J Am Coll Radiol 2016;13(10):1210–1214.
3. Yu J-PJ, Kansagra AP, Mongan J. The radiologist's workflow environment: evaluation of disruptors and potential implications. J Am Coll Radiol 2014;11(6):589–593.
4. Balint BJ, Steenburg SD, Lin H, Shen C, Steele JL, Gunderman RB. Do telephone call interruptions have an impact on radiology resident diagnostic accuracy? Acad Radiol 2014;21(12):1623–1628.
5. Liles AL, Francis IR, Kalia V, Kim J, Davenport MS. Common Causes of Outpatient CT and MRI Callback Examinations: Opportunities for Improvement. AJR Am J Roentgenol 2020;214(3):487–492.
6. Gyftopoulos S, Kim D, Aaltonen E, Horwitz LI. Patient Recall Imaging in the Ambulatory Setting. AJR Am J Roentgenol 2016;206(4):787–791.
7. Raj A, Jindal R, Singh AK, Pal A. A Study of Recent Advancements in Deep Learning for Natural Language Processing. In: 2023 IEEE World Conference on Applied Intelligence and Computing (AIC). IEEE, 2023; 300–306.
8. Casey A, Davidson E, Poon M, et al. A systematic review of natural language processing applied to radiology reports. BMC Med Inform Decis Mak 2021;21(1):179.
9. Linna N, Kahn CE Jr. Applications of natural language processing in radiology: A systematic review. Int J Med Inform 2022;163:104779.
10. Brown AD, Marotta TR. A Natural Language Processing-based Model to Automate MRI Brain Protocol Selection and Prioritization. Acad Radiol 2017;24(2):160–166.
11. Chillakuru YR, Munjal S, Laguna B, et al. Development and web deployment of an automated neuroradiology MRI protocoling tool with natural language processing. BMC Med Inform Decis Mak 2021;21(1):213.
12. Kalra A, Chakraborty A, Fine B, Reicher J. Machine Learning for Automation of Radiology Protocols for Quality and Efficiency Improvement. J Am Coll Radiol 2020;17(9):1149–1158.
13. López-Úbeda P, Díaz-Galiano MC, Martín-Noguerol T, Luna A, Ureña-López LA, Martín-Valdivia MT. Automatic medical protocol classification using machine learning approaches. Comput Methods Programs Biomed 2021;200:105939.
14. Eghbali N, Siegal D, Klochko C, Ghassemi MM. Automation of Protocoling Advanced MSK Examinations Using Natural Language Processing Techniques. AMIA Jt Summits Transl Sci Proc 2023;2023:118–127.
15. Lau W, Aaltonen L, Gunn M, Yetisgen M. Automatic Assignment of Radiology Examination Protocols Using Pre-trained Language Models with Knowledge Distillation. AMIA Annu Symp Proc 2022;2021:668–676.
16. Gertz RJ, Bunck AC, Lennartz S, et al. GPT-4 for Automated Determination of Radiological Study and Protocol based on Radiology Request Forms: A Feasibility Study. Radiology 2023;307(5):e230877.
17. Wong KA, Hatef A, Ryu JL, Nguyen XV, Makary MS, Prevedello LM. An Artificial Intelligence Tool for Clinical Decision Support and Protocol Selection for Brain MRI. AJNR Am J Neuroradiol 2023;44(1):11–16.
18. Xavier BA, Chen PH. Natural Language Processing for Imaging Protocol Assignment: Machine Learning for Multiclass Classification of Abdominal CT Protocols Using Indication Text Data. J Digit Imaging 2022;35(5):1120–1130.
19. Trivedi H, Mesterhazy J, Laguna B, Vu T, Sohn JH. Automatic Determination of the Need for Intravenous Contrast in Musculoskeletal MRI Examinations Using IBM Watson's Natural Language Processing Algorithm. J Digit Imaging 2018;31(2):245–251.
20. Lee YH. Efficiency Improvement in a Busy Radiology Practice: Determination of Musculoskeletal Magnetic Resonance Imaging Protocol Using Deep-Learning Convolutional Neural Networks. J Digit Imaging 2018;31(5):604–610.
21. Nencka AS, Sherafati M, Goebel T, Tolat P, Koch KM. Deep-learning based Tools for Automated Protocol Definition of Advanced Diagnostic Imaging Exams. ArXiv 2106.08963 [preprint] https://arxiv.org/abs/2106.08963. Posted May 28, 2021. Accessed November 27, 2023.

22. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol 1 (Long and Short Papers). Association for Computational Linguistics, 2019; 4171–4186.
23. Yan A, McAuley J, Lu X, et al. RadBERT: Adapting Transformer-based Language Models to Radiology. Radiol Artif Intell 2022;4(4):e210258.
24. Brown TB, Mann B, Ryder N, et al. Language Models are Few-Shot Learners. ArXiv 2005.14165 [preprint] https://arxiv.org/abs/2005.14165. Posted May 28, 2020. Accessed November 27, 2023.
25. Elyoseph Z, Hadar-Shoval D, Asraf K, Lvovsky M. ChatGPT outperforms humans in emotional awareness evaluations. Front Psychol 2023;14:1199058.
26. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198.
27. Virtanen A, Kanerva J, Ilo R, et al. Multilingual is not enough: BERT for Finnish. ArXiv 1912.07076 [preprint] https://arxiv.org/abs/1912.07076. Posted December 15, 2019. Accessed February 14, 2021.
28. Happonen T, Nyman M, Ylikotila P, Mattila K, Hirvonen J. Imaging Outcomes of Emergency MR Imaging in Dizziness and Vertigo: A Retrospective Cohort Study. AJNR Am J Neuroradiol 2024;45(6):819–825.
29. Happonen T, Nyman M, Ylikotila P, Merisaari H, Mattila K, Hirvonen J. Diagnostic yield of emergency MRI in non-traumatic headache. Neuroradiology 2023;65(1):89–96.
30. GPT-3.5 Turbo. OpenAI. https://platform.openai.com/docs/models. Accessed March 2024.