Title

Application of a Domain-specific BERT for Detection of Speech Recognition Errors in Radiology Reports.

Permalink

https://escholarship.org/uc/item/86k0t1xf

Journal

Radiology. Artificial intelligence, 4(4)

ISSN

2638-6100

Authors

Chaudhari, Gunvant R
Liu, Tengxiao
Chen, Timothy L
et al.

Publication Date

2022-07-01

DOI

10.1148/ryai.210185

Peer reviewed

Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Application of a Domain-specific BERT for Detection of Speech Recognition Errors in Radiology Reports

Gunvant R. Chaudhari, BS • Tengxiao Liu • Timothy L. Chen, MD • Gabby B. Joseph, PhD • Maya Vella, MD • Yoo Jin Lee, MD • Thienkhai H. Vu, MD, PhD • Youngho Seo, PhD • Andreas M. Rauschecker, MD, PhD • Charles E. McCulloch, PhD • Jae Ho Sohn, MD, MS

From the Department of Radiology and Biomedical Imaging (G.R.C., T.L., T.L.C., G.B.J., M.V., Y.J.L., T.H.V., Y.S., A.M.R., J.H.S.) and Department of Epidemiology and Statistics (C.E.M.), University of California San Francisco, 505 Parnassus Ave, San Francisco, CA 94143. Received July 5, 2021; revision requested August 11; revision received April 11, 2022; accepted May 10. **Address correspondence to** J.H.S. (email: *sohn87@gmail.com*).

Authors declared no funding for this work.

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

See also the commentary by Abajian and Cheung in this issue.

Radiology: Artificial Intelligence 2022; 4(4):e210185 • **https://doi.org/10.1148/ryai.210185** • **Content code:** AI

Purpose: To develop radiology domain–specific bidirectional encoder representations from transformers (BERT) models that can identify speech recognition (SR) errors and suggest corrections in radiology reports.

Materials and Methods: A pretrained BERT model, Clinical BioBERT, was further pretrained on a corpus of 114 008 radiology reports between April 2016 and August 2019 that were retrospectively collected from two hospitals. Next, the model was fine-tuned on a training dataset of generated insertion, deletion, and substitution errors, creating Radiology BERT. This model was retrospectively evaluated on an independent dataset of radiology reports with generated errors (*n* = 18 885) and on unaltered report sentences (*n* = 2000) and prospectively evaluated on true clinical SR errors (*n* = 92). Correction Radiology BERT was separately trained to suggest corrections for detected deletion and substitution errors. Area under the receiver operating characteristic curve (AUC) and bootstrapped 95% CIs were calculated for each evaluation dataset.

Results: Radiology-specific BERT had AUC values of >.99 (95% CI: >0.99, >0.99), 0.94 (95% CI: 0.93, 0.94), 0.98 (95% CI: 0.98, 0.98), and 0.97 (95% CI: 0.97, 0.97) for detecting insertion, deletion, substitution, and all errors, respectively, on the independently generated test set. Testing on unaltered report impressions revealed a sensitivity of 82% (28 of 34; 95% CI: 70%, 93%) and specificity of 88% (1521 of 1728; 95% CI: 87%, 90%). Testing on prospective SR errors showed an accuracy of 75% (69 of 92; 95% CI: 65%, 83%). Finally, the correct word was the top suggestion for 45.6% (475 of 1041; 95% CI: 42.5%, 49.3%) of errors.

Conclusion: Radiology-specific BERT models fine-tuned on generated errors were able to identify SR errors in radiology reports and suggest corrections.

Supplemental material is available for this article.

Computer-aided speech recognition (SR) has been widely adopted by radiology departments nationwide, with 85% of practices using it nationwide in 2018 (1). Continual advances in hardware and software have increased accuracy of SR systems (2). However, SR software remains imperfect and regularly produces errors that can alter clinical meaning and interpretation (3,4). Compared with the errors found in typed reports, errors from SR software are rarely spelling mistakes and more commonly semantic or grammatical errors. Traditional document error–checking algorithms are not well suited to detect semantic errors (5). Several approaches, including use of co-occurrence relations (6), image metadata (7), and neural sequence to sequence models (8), have been previously explored to identify SR errors in radiology reports.

Recent use of transformer-based architectures in natural language processing (NLP), such as the bidirectional encoder representations from transformers (BERT), has resulted in substantial improvement in benchmark NLP tasks compared with previous architectures (9). BERT models pretrained on large text corpora have been released for open use (9). Studies have shown that these BERT models can be further pretrained or fine-tuned with small datasets for specific downstream NLP tasks (10,11). In radiology, BERT has been used to classify knee osteoarthritis reports, extract spatial relation information, identify significant findings in chest radiograph reports, extract ischemic stroke characteristics, and identify communication urgency (12–17). However, a comprehensive model for detecting dictation errors in radiology reports across multiple imaging modalities has yet to be established, to our knowledge.

In this study, we applied BERT to create a robust context-based tool for handling radiology report dictation errors. We hypothesized that we could use transfer learning from a pretrained medicine-specific BERT model to create a radiology-specific BERT model, and that this model could then be fine-tuned to automatically detect report errors and suggest corrections at the token level.

Abbreviations

AUC = area under the receiver operating characteristic curve, BERT = bidirectional encoder representations from transformers, NLP = natural language processing, SR = speech recognition

Summary

A pretrained bidirectional encoder representations from transformers (BERT) model that has been adapted to a radiology corpus and fine-tuned to identify speech recognition errors in radiology reports was evaluated using retrospective and prospective analyses.

Key Points

- A radiology-specific bidirectional encoder representations from transformers (BERT) model fine-tuned for report error detection identified insertion, deletion, and substitution errors with area under the curve (AUC) values of >0.99, 0.94, and 0.98, respectively, on a generated errors dataset.
- Testing on errors in retrospectively collected signed radiology reports showed an AUC of 0.95 with sensitivity of 82% and specificity of 99%.
- Testing the model on real-time, prospectively collected speech recognition errors from clinical workflow demonstrated an AUC of 0.88 and sentence-wise accuracy of 75%.

Keywords

Computer Applications, Technology Assessment

Materials and Methods

Datasets and Corpora

This retrospective model development study was approved by the human ethics board of our institution and was conducted in accordance with the Helsinki Declaration of 1975, as revised in 2013, with consent waived. A total of 121 396 radiology reports, each with a unique accession number, were used for model training. This training corpora was aggregated from two medical institutions, which partially share staff members, during three separate time periods: 5295 CT reports between March 2019 and April 2019 from University of California San Francisco (UCSF); 38 222 CT, MRI, PET, and US reports between January 2017 and March 2017 from UCSF; and 77 879 reports containing radiography, CT, US, MRI, mammography, procedural, and nuclear medicine studies between April 2016 and September 2016 from Zuckerberg San Francisco General Hospital. All reports were stripped of patient-identifying labels to the best of our ability, and 7388 reports were excluded from the dataset due to duplicates (n = 933), external studies that lacked a radiology report (n = 4870), or nonstandard reports (n = 1585) with formatting deviating from institutional standard that prevented algorithmic impression extraction (Fig 1). Next, impression section texts were extracted from the reports and segmented into sentences with spaCy *(https://spacy.io)* and Python 3.6, resulting in a dataset of 114 008 reports, 470 157 sentences, 4 758 081 words, and 7 354 058 tokens. Only the impression section was used for all training and testing because it is the most expressive part of the report, allowed for the most consistent segmentation, and trained the most stable model during the exploratory phase of model search.

Separately, we extracted an independent test dataset consisting of 18 885 reports of comprehensively and consecutively collected studies from August 16, 2020, to August 29, 2020, from UCSF. The same preprocessing steps as for the training dataset were applied to yield an independently generated test set with 13 928 reports (n = 48 592 sentences, 657 152 words, 940 211 tokens). A subset of 2000 unaltered sentences (n = 21 367 words, 30 982 tokens) from this independent test dataset was randomly sampled to manually analyze model accuracy.

A prospective clinical test dataset was created by four radiologists (J.H.S., Y.J.L., M.V.) from UCSF over the course of 28 days in December 2020 and March 2021. Whenever SR produced a sentence that contained an error, the errored sentence was manually marked and added to this dataset in real time (n = 92 sentences, 1358 words, 2006 tokens).

All datasets contained reports that had been dictated with PowerScribe (Nuance Communications; version 2016–2019).

Data Preprocessing and Preparation: Automated Error Generator

To fine-tune our BERT models for detecting SR errors, we simulated errors that are likely to occur in dictated reports. We designed a metaphone-based error generator that creates three types of errors: deletion, insertion, and substitution (Appendix E1 [supplement]). From the results of a previous study analyzing clinical SR errors (18), the probability of a given word being changed into an error was set to 7.4%, and the relative proportion of insertion, deletion, and substitution errors was set to 0.347, 0.270, and 0.383, respectively.

Model Predevelopment: Creating a Radiology-specific BERT Model through Additional Pretraining

To create a radiology-specific BERT model, we initialized our model with parameters from Clinical BioBERT (10,19). We further trained these models on both the masked language model and next sentence prediction tasks (9) using the training corpora dataset and similar hyperparameters (Appendix E1 [supplement]) to the training of Clinical BioBERT (19). For the task of error correction, a separate but analogous BERT model was trained from Clinical BioBERT on solely the masked language model task with the training corpora dataset (Appendix E1 [supplement]) to create Correction Radiology BERT.

Model Development: Fine-Tuning BERT to Detect Report Errors

We devised a token classification task to detect single-token errors in radiology reports. Each input token was labeled as a normal token, an insertion error, a deletion error, or a substitution error. For insertion and substitution errors, the model was trained to flag the suspected errored word; for deletions, it was trained to flag the word after the suspected deletion. A fully connected linear layer for token classification with softmax output was added on top of the BERT hidden states output (Fig 2).

The training corpora underwent processing by the automated error generator to create an "errored" training corpora

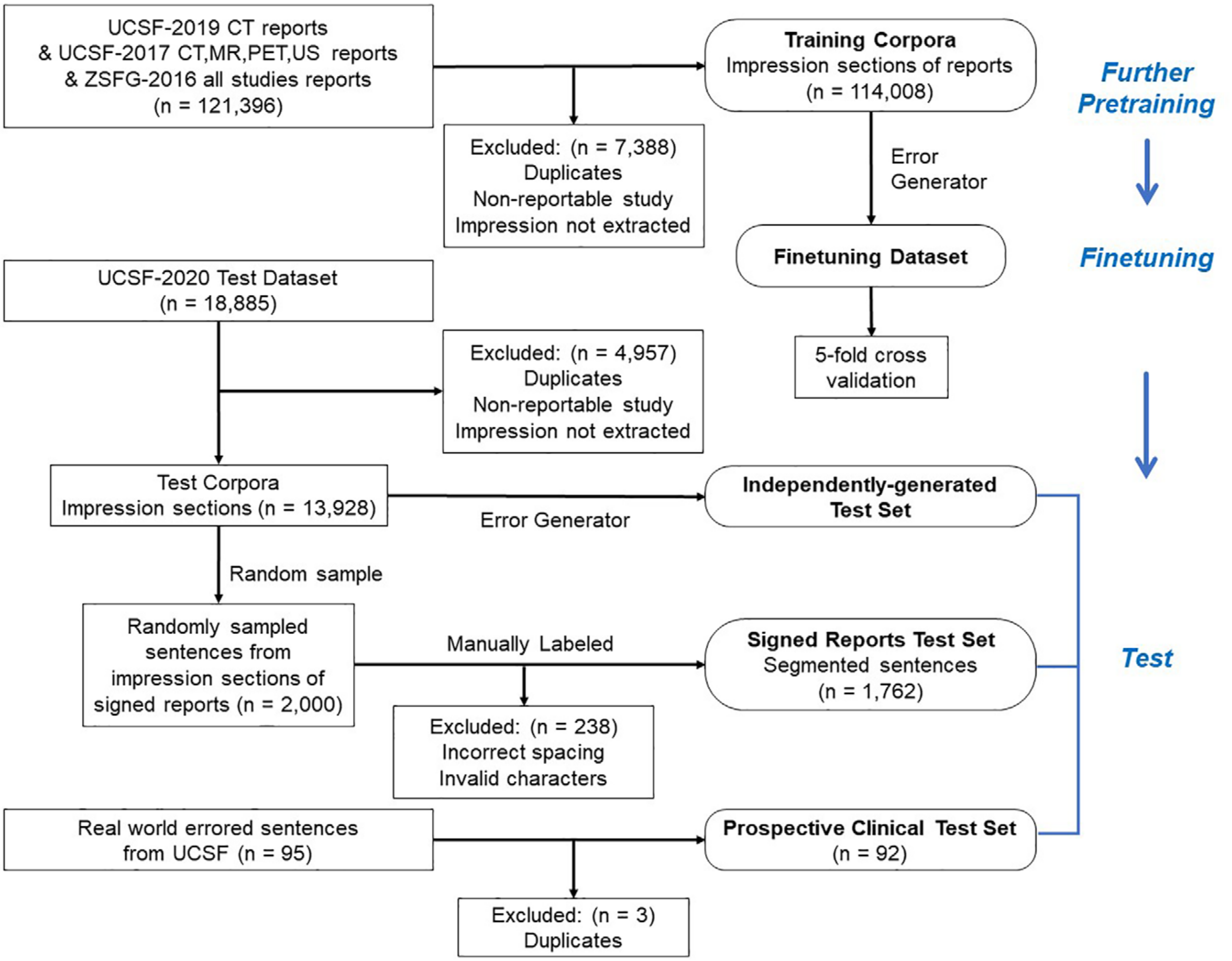


Figure 1: Inclusion and exclusion criteria of text data for pretraining, fine-tuning, and test sets. Training corpora were used in further pretraining and then corrupted through the error generator for model fine-tuning. The independently generated test set, signed reports test dataset, and prospective clinical dataset were all used in evaluation. UCSF = University of California San Francisco, ZSFG = Zuckerberg San Francisco General Hospital.

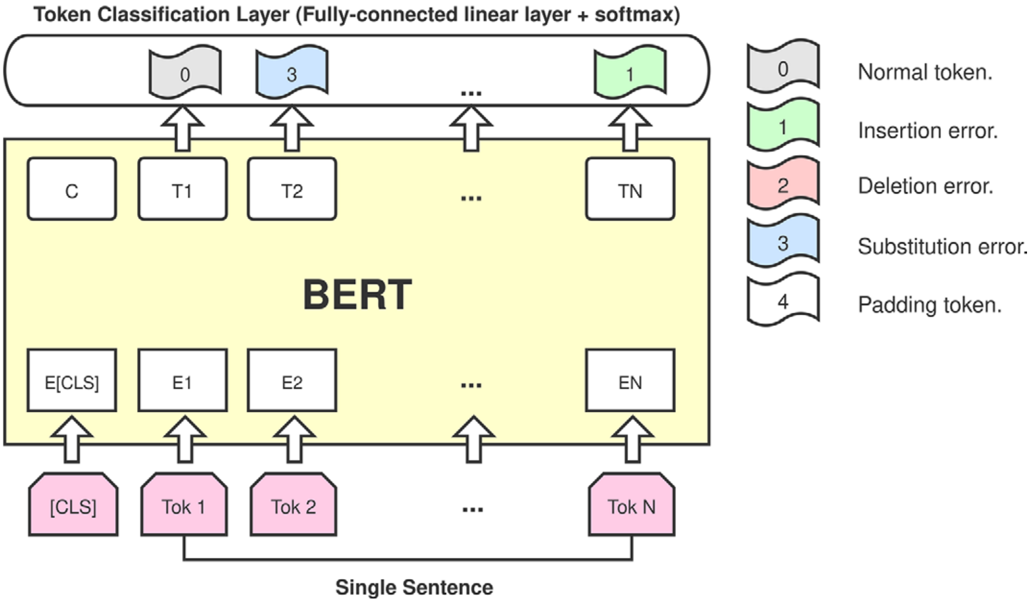
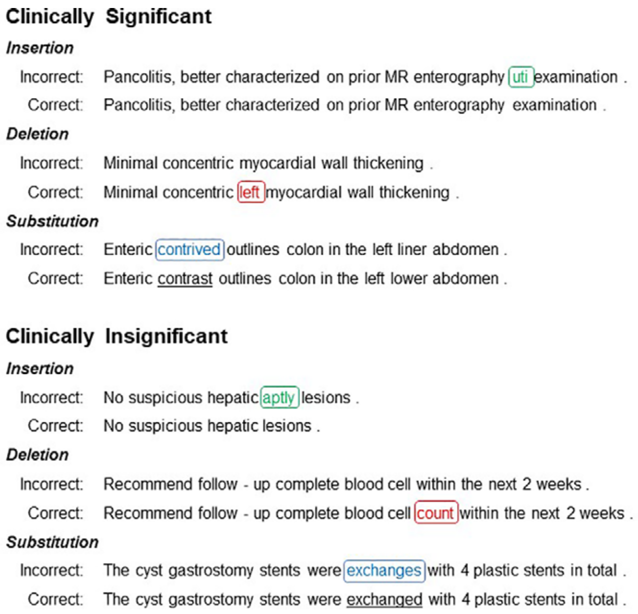


Figure 2: Fine-tuning network depiction. The model was fine-tuned with automatically generated error data. Input tokens ("Tok1,""Tok2,"...) are fed into bidirectional encoder representations from transformers (BERT) as embeddings ("E1,""E2,"...) with a special token [CLS] indicating the start of a sentence. Output ("T1,""T2,"...) from the BERT model was fed into a fully connected linear classification layer with softmax activation. The classification layer generated five labels (0–4) that denote the input token as a normal token, an insertion error, a deletion error, a substitution error, and a padding token, respectively. C represents the unused class label for the input sentence.



Clinically Significant

Insertion

Incorrect: Pancolitis, better characterized on prior MR enterography uti examination .

Correct: Pancolitis, better characterized on prior MR enterography examination .

Deletion

Incorrect: Minimal concentric myocardial wall thickening .

Correct: Minimal concentric left myocardial wall thickening .

Substitution

Incorrect: Enteric contrived outlines colon in the left liner abdomen .

Correct: Enteric contrast outlines colon in the left lower abdomen .

Clinically Insignificant

Insertion

Incorrect: No suspicious hepatic aptly lesions .

Correct: No suspicious hepatic lesions .

Deletion

Incorrect: Recommend follow - up complete blood cell within the next 2 weeks .

Correct: Recommend follow - up complete blood cell count within the next 2 weeks .

Substitution

Incorrect: The cyst gastrostomy stents were exchanges with 4 plastic stents in total .

Correct: The cyst gastrostomy stents were exchanged with 4 plastic stents in total .

Figure 3: Examples of sentences automatically errored with insertions (green), deletions (red), and substitutions (blue) that would or would not affect clinical significance according to consensus of three authors (J.H.S., T.L.C., G.R.C.). Samples were deemed not clinically significant if the original meaning was able to be reasonably derived given the errored sentence alone and if downstream management would not change.

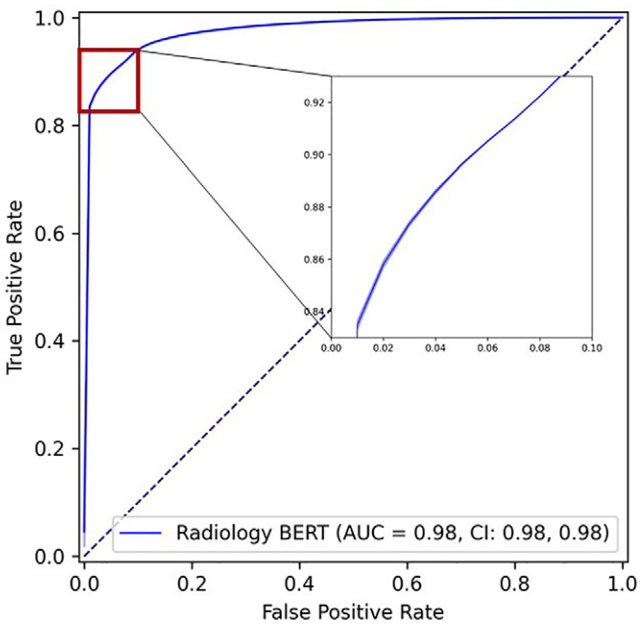


Figure 4: Receiver operating characteristic (ROC) curve depicting the fine-tuned performance of the Radiology bidirectional encoder representations from transformers (BERT) model. Analyses were performed on the holdout validation sets, and results for the all-errors class are shown. The shading shows 1 SD of the ROC curve, and the 95% CI is reported. AUC = area under the curve.

that was then used to train the model. We named the trained model, now fine-tuned to a classification task, Radiology BERT. PyTorch (version 1.6.0) and the HuggingFace transformers library (version 3.4.0) were used to implement these methods (20).

Model Evaluation

We evaluated our model using three tasks. First, we determined performance on automatically generated errors using the holdout validation sets (n = 114 008 reports) and independently generated test set (n = 18 885 reports). The model was trained and tested on errored impression phrases. To evaluate the effectiveness of the error generator algorithm, two medical trainees (G.R.C., T.L.C.) separately analyzed 509 errored sentences to determine if they had clinically significant errors that would affect clinical interpretation, as defined in Alsentzer et al (19).

Second, we determined the performance of the model on signed reports from the independent test set. A medical trainee (G.R.C.) manually analyzed 2000 randomly selected impression sentences and marked any errors. Some sentences (n = 238) were excluded for incorrect sentence segmentation due to incorrect spacing or the presence of invalid characters that would not be encountered in a real-world radiology workflow (Fig 1).

Finally, we evaluated the performance of our approach on true SR errors collected prospectively in a real clinical workflow. Errored sentences that appeared during report dictation were collected in real time before they were corrected. All errored sentences were corrected and categorized by a board-eligible radiologist (J.H.S.) and two medical trainees (G.R.C., T.L.C.) (Fig 3).

Statistical Analysis

During optimal model search, each experiment was run with fivefold cross-validation for both pretraining and fine-tuning data. The all-errors class is defined as one minus the value of the normal class. For all sentence-level analyses, one minus each sentence's lowest value for the normal class output was used as the likelihood that the sentence has an error. For this analysis, metrics were calculated at the optimal threshold (point on receiver operating characteristic curve closest to [0,1]). For the prospective clinical dataset's sentence-level metrics, the same threshold as for the signed reports test set was used because the prospective dataset lacked negative samples. The 95% CIs were generated using bootstrapping with resampling at the report or sentence level to accommodate clustering. All statistical analyses were performed in Python 3.6 using the scikit-learn package, and a P value of less than .05 was considered to indicate a significant difference. Additional details are provided in Appendix E1 (supplement).

Results

Dataset Characteristics

Generated errors from 114 008 radiology reports were used for training, and those from 18 885 reports for initial testing. The most represented body parts from these two datasets were abdomen-pelvis (25% of reports) and chest (24% of reports) (Table 1). Both datasets covered seven modalities, of which radiography and CT make up more than 60% of the total studies. A total of 470 157 sentences from the training set were used to generate automatically corrupted sentences for fine-tuning. Assessment of a subset of these corrupted sentences by

Table 1: Characteristics of Training, Validation, and Individual Test Sets Prior to Exclusion

Characteristic	Training Corpora* (*n* = 121 396)	Hospital 1 2020 Test Dataset (*n* = 18 885)	Prospective Clinical Test Dataset (*n* = 95)
Sex			
Male	21 043 (48)†	8517 (45)	54 (57)
Female	22 460 (52)†	10 355 (55)	41 (43)
Other‡	14 (0)†	13 (0)	0 (0)
Study type			
Radiography	44 658 (37)	7757 (41)	26 (27)
CT	33 576 (28)	4005 (21)	61 (64)
US	19 731 (16)	2293 (12)	0 (0)
MRI	13 632 (11)	2886 (15)	8 (8)
Mammography	5809 (5)	690 (4)	0 (0)
Procedure	2143 (2)	494 (3)	0 (0)
Nuclear medicine	1847 (2)	760 (4)	0 (0)
Body part imaged			
Head	13 730 (11)	1898 (10)	0 (0)
Neck	4347 (4)	769 (4)	0 (0)
Chest	29 875 (25)	4480 (24)	95 (100)
Breast	7276 (6)	990 (5)	0 (0)
Abdomen/pelvis	30 385 (25)	4723 (25)	0 (0)
Spine	10 029 (8)	2022 (11)	0 (0)
Extremity	21 887 (18)	2742 (15)	0 (0)
Whole body	2191 (2)	689 (4)	0 (0)
Other	1676 (1)	572 (3)	0 (0)
Radiologists represented	152§	105	4

Note.—Data presented as numbers of reports with percentages in parentheses.
* Training corpora refers to aggregated reports from hospital 1 2019, hospital 1 2017, and hospital 2 2016.
† Sex information was not available in hospital 2 2016 dataset.
‡ Patient self-reported "Other" sex category in the medical record.
§ Signing radiologist information was not available in hospital 1 2017 dataset.

two medical trainees showed that 31.0% (158 of 509) of errored sentences generated using our algorithm would change clinical interpretation. There was moderate agreement between the readers (Cohen κ: 0.435; 95% CI: 0.357, 0.514; Table E1 [supplement]).

Dictation Error Detection Model Evaluation

The fine-tuned Radiology BERT showed area under the receiver operating characteristic curve (AUC) values of >0.99 (95% CI: >0.99, >0.99), 0.96 (95% CI: 0.96, 0.96), 0.99 (95% CI: 0.99, 0.99), and 0.98 (95% CI: 0.98, 0.98) (Fig 4) for insertion, deletion, substitution, and all errors, respectively, on the amalgamated holdout validation sets from five-fold cross-validation (Table 2). On generated errors from the independent test set, Radiology BERT had a token-level all-error AUC of 0.97 (95% CI: 0.97, 0.97) and a sentence-level all-error AUC of 0.96 (95% CI: 0.96, 0.96). Evaluation of Radiology BERT on the signed reports test set, a dataset of 2000 unaltered sentences with clinical dictation errors inadvertently signed into the record, revealed AUCs of 0.72 (95% CI: 0.49, >0.99), 0.87 (95% CI: 0.71, >0.99), 0.95 (95% CI: 0.86, >0.99), and 0.95 (95% CI: 0.89, 0.99) for insertion, deletion, substitution, and all errors, respectively (Table 3). Furthermore, Radiology BERT had an AUC of 0.89 (0.83, 0.94) for detecting whether a given report sentence contained an error, which corresponded with a sentence-level sensitivity of 82% (28 of 34, 95% CI: 70%, 93%) and specificity of 88% (1521 of 1728, 95% CI: 87%, 90%). Most errors in the signed reports test set were deletion errors (19 of 34, 56%), while insertion errors were rare (two of 34, 6%).

On prospectively collected dictation errors, Radiology BERT had AUCs of 0.77 (95% CI: 0.58, 0.99), 0.61 (95% CI: 0.42, 0.86), 0.88 (95% CI: 0.84, 0.92), and 0.88 (95% CI: 0.84, 0.92) for insertion, deletion, substitution, and all errors, respectively (Table 3). This performance corresponded to an error recognition accuracy of 75% (69 of 92; 95% CI: 65%, 83%) at the sentence level. An analysis of these collected error sentences by a board-certified radiologist (J.H.S.) and a medical trainee (G.R.C.) revealed that 10 sentences (11% of dataset) were grammatically and medically correct and could not be labeled as

Table 2: Metrics of Radiology BERT Performance on the Holdout Validation Sets and Independently Generated Test Set

Set	AUC	Sensitivity	Specificity	PPV	NPV
Holdout validation sets*					
Insertion (*n* = 177 870 tokens)	>0.99 (> 0.99, > 0.99)	95.6% (95.5%, 95.7%)	99.9% (99.8%, 99.9%)	94.1% (93.9%, 94.2%)	99.9% (99.9%, 99.9%)
Deletion (*n* = 144 636 tokens)	0.96 (0.96, 0.96)	62.3% (61.9%, 62.8%)	99.7% (99.7%, 99.7%)	81.5% (81.1%, 81.8%)	99.3% (99.2%, 99.3%)
Substitution (*n* = 240 943 tokens)	0.99 (0.99, 0.99)	80.2% (79.9%, 80.4%)	99.8% (99.8%, 99.8%)	92.0% (91.8%, 92.2%)	99.3% (99.3%, 99.3%)
All errors (*n* = 563 449 errored tokens)	0.98 (0.98, 0.98)	83.2% (83.0%, 83.4%)	99.5% (99.5%, 99.5%)	93.5% (93.4%, 93.7%)	98.6% (98.6%, 98.6%)
All errors (*n* = 219 940 errored sentences)	0.98 (0.98, 0.98)	93.6% (93.5%, 93.7%)	95.7% (95.6%, 95.8%)	96.5% (96.4%, 96.5%)	92.3% (92.2%, 92.5%)
Independently generated test set†					
Insertion (*n* = 21 982 tokens)	>0.99 (> 0.99, > 0.99)	94.8% (94.5%, 95.1%)	99.7% (99.7%, 99.8%)	89.7% (89.1%, 90.3%)	99.9% (99.9%, 99.9%)
Deletion (*n* = 18 087 tokens)	0.94 (0.93, 0.94)	52.4% (51.2%, 53.6%)	99.4% (99.4%, 99.5%)	65.1% (63.9%, 66.4%)	99.1% (99.0%, 99.1%)
Substitution (*n* = 29 417 tokens)	0.98 (0.98, 0.98)	75.0% (74.1%, 75.8%)	99.5% (99.5%, 99.5%)	82.7% (81.9%, 83.5%)	99.2% (99.2%, 99.2%)
All errors (*n* = 69 486 errored tokens)	0.97 (0.97, 0.97)	78.9% (78.4%, 79.4%)	98.9% (98.9%, 98.9%)	85.0% (84.6%, 85.5%)	98.3% (98.3%, 98.4%)
All errors (*n* = 27 251 errored sentences)	0.96 (0.96, 0.96)	89.4% (89.0%, 89.8%)	92.4% (92.0%, 92.8%)	93.8% (93.4%, 94.1%)	87.3% (86.8%, 87.7%)

Note.—Values are presented with 95% CIs in parentheses. AUC = area under the curve, BERT = bidirectional encoder representations from transformers, NPV = negative predictive value, PPV = positive predictive value.

* Holdout validation set (amalgamated): *n* = 7 354 058 tokens, 470 157 sentences.

† Independently generated test set: *n* = 30 982 tokens, 1786 sentences.

human error without context of the imaging and entire report, which the model does not have access to. Examples of model performance on sentences from all testing datasets are provided in Figure E2A–E2C (supplement).

Dictation Error Correction Model Evaluation

For word candidate prediction, we evaluated the separate trained model, Correction Radiology BERT, on 1041 sentences sampled from the independent test set. This model was able to identify the correct word as the top suggestion for 45.6% (475 of 1041; 95% CI: 42.5%, 49.3%) (Table E5 [supplement]) of substitution and deletion errors, and 55.9% (582 of 1041; 95% CI: 52.9%, 59.0%) of errors had the correct word within the top three suggestions. Examples of correction model performance are provided in Figure E2D (supplement).

Error Analysis

Table 4 demonstrates representative examples of incorrect predictions by the model and suspected reasons for the errors. These represent cases from the retrospective evaluation of final signed reports and included both false-positive and false-negative cases. Out of 2000 sentences, 34 were identified to have a true error. Most false-positive predictions by the model were deletion errors (88 of 128, 69%), and false-negative predictions consisted largely of deletion (five of 10, 50%) and substitution (four of 10, 40%) errors.

Discussion

We have shown that a radiology domain–specific BERT model can effectively flag potential errors in radiology reports generated by SR and provide correction suggestions. Our best-performing model, Radiology BERT, was further pretrained from Clinical BioBERT and fine-tuned on an automatically generated errored corpus. It achieved average AUCs of >0.99, 0.94, 0.98, and 0.97 for insertion, deletion, substitution, and all errors, respectively, on the independently generated test dataset. Additionally, evaluation on the retrospective signed reports test dataset and prospective clinical dataset demonstrated AUCs of 0.95 and 0.88, respectively.

Prior work for detecting errors in SR reports used seq2seq and involved training models using single body-part and modality data (8). Numerical comparison between this approach and our proposed BERT approach is not possible due to the lack of open code and data. However, our training data consisted of a multitude of body parts, modalities, and sequences to create a model with broader applicability. Furthermore, large pretrained transformer-based models such as BERT are known to be more capable of natural language understanding

Table 3: Metrics of Radiology BERT Performance on the Signed Reports Test Dataset and Prospective Clinical Dataset

Dataset	AUC	Sensitivity	Specificity	PPV	NPV
Signed reports test dataset*					
Insertion (*n* = 2 tokens)	0.72 (0.49, > 0.99)	50% (0%, 100%)	100% (100%, 100%)	5% (0.0%, 19%)	100% (100%, 100%)
Deletion (*n* = 16 tokens)	0.87 (0.71, > 0.99)	75% (50%, 100%)	99% (99%, 99%)	5% (2%, 8%)	100% (100%, 100%)
Substitution (*n* = 44 tokens)	0.95 (0.86, > 0.99)	87% (69%, 98%)	100% (100%, 100%)	45% (30%, 62%)	100% (100%, 100%)
All errors (*n* = 62 errored tokens)	0.95 (0.89, 0.99)	82% (68%, 94%)	99% (99%, 99%)	15% (9%, 20%)	100% (100%, 100%)
All errors (*n* = 34 errored sentences)	0.89 (0.83, 0.94)	82% (70%, 93%)	88% (87%, 90%)	13% (9%, 18%)	100% (99%, 100%)
Prospective clinical dataset†					
Insertion (*n* = 9 tokens)	0.77 (0.58, 0.99)	23% (0%, 60%)	99% (99%, 100%)	13% (0%, 33%)	100% (99%, 100%)
Deletion (*n* = 10 tokens)	0.61 (0.42, 0.86)	21% (0%, 60%)	99% (98%, 99%)	8% (0%, 19%)	100% (99%, 100%)
Substitution (*n* = 206 tokens)	0.88 (0.84, 0.92)	36% (26%, 47%)	100% (99%, 100%)	90% (80%, 99%)	93% (91%, 95%)
All errors (*n* = 225 errored tokens)	0.88 (0.84, 0.92)	44% (35%, 54%)	99% (98%, 99%)	81% (72%, 90%)	93% (91%, 95%)
All errors (*n* = 92 errored sentences)	NA	75% (65%, 83%)	NA	NA	NA

Note.—Values are presented with 95% CIs in parentheses. AUC, specificity, PPV, and NPV for the prospective clinical dataset could not be calculated at the sentence level because all collected sentences were errored. AUC = area under the curve, BERT = bidirectional encoder representations from transformers, NA = not available, NPV = negative predictive value, PPV = positive predictive value.

* Signed reports test dataset: *n* = 1762 sentences, 30 982 tokens.

† Prospective clinical dataset: *n* = 92 sentences, 2006 tokens.

Table 4: Radiology BERT Error Analysis

No.	Sentence	Flagged Error Type	True Error	Suspected Reason
False-positive 1	No evidence of {{screw}} fracture or failure.	Insertion	None	"No evidence of fracture" is a valid phrase that was probably much more common in corpus.
False-positive 2	Soft tissues {{unremarkable}}.	Deletion	None	Model was trained on many sentences of "Soft tissues *are* unremarkable."
False-positive 3	Mildly dilated small bowel segments with {{ample}} gas in the colon.	Substitution	None	"Ample" is probably not a commonly used word in radiology reports
False-negative 1	An enhancing exophytic {{enhancing}} mass at the left superior renal pole is suspicious for renal cell carcinoma.	None	Insertion "enhancing"	Repeated word insertion errors are rare in the automatically generated error training dataset
False-negative 2	Lines/drains/medical devices: Feeding tube has been slightly {{advanced the}} tip pointing at the gastric outlet.	None	Deletion "advanced with the"	Inconsistent grammar usage in radiology reports
False-negative 3	Scattered interphalangeal joint arthrosis most pronounced {{an}} moderate at the second DIP joint.	None	Substitution "and" to "an"	Uncorrected "an" error may be common in training radiology reports

Note.—Incorrect classifications made by Radiology BERT on signed reports test dataset. Words in question are encased in double braces.

tasks than a seq2seq approach (21), largely due to the transformers' use of bidirectional context for each token. Furthermore, we chose to train on publicly released BERT models trained on large corpora to improve our model performance and generalizability.

The clinical issue addressed in this study is that errors in radiology reports are pervasive problems that decrease clinician and patient satisfaction toward radiology and may affect patient care. Traditional spell checkers cannot identify most SR errors in radiology reports because they only recognize spelling and grammar errors. Our error detection and correction approach is intended to add value to daily clinical routine by reviewing radiologists' dictated reports at the time of signing and flagging any potential unusual, inappropriate, or out-of-context words for radiologists to review. This will reduce the burden on radiologists by reducing the frequency of providing necessary addendums or corrections to reports. The reduction of dictation errors can improve communication and trust between the radiologist and readers of the radiology reports.

We evaluated our model on three different test datasets, each of which served a unique purpose. The independently generated test dataset was used to verify that a high-performing BERT model was successfully trained and could perform well on a large permutation of errors that could theoretically be found in radiology reports. The signed report test dataset contains text that has already undergone proofreading and provides insight into how the model may perform in the use-case of checking a report before signing. Finally, the prospective clinical test set evaluated the algorithm's ability to detect SR errors that appeared during dictation prior to any proofreading, which is the algorithm's intended use case.

Sentence fragments are often used in radiology reports. It is worth noting that acceptable syntax is variable across study sites and even between radiologists. This inconsistency may lead to false-positive findings as the model may be trained on one particular syntax but be presented with an alternative, but acceptable, syntax (eg, "Soft tissues unremarkable," Table 4). Other errors such as negation, laterality, or insertion or deletion of "no" are generally impossible to detect at sentence level because the sentence is syntactically, grammatically, and medically correct. To identify these mistakes, radiologists often need to draw on additional evidence, such as the imaging from the study. Several such errors decreased our model's performance on the prospective clinical test set.

As shown in Table 4, we analyzed some errors from the retrospective evaluation of final signed reports. In false-positive cases 1 and 3, the model flagged the sentences likely because they deviated from some common phrases in the training corpus. False-negative findings exposed the vulnerability of the automatically generated corpus created from inherently imperfect reports. The false-negative case 1 is likely due to the rarity of repeated word insertion cases in our training dataset, which likely made the model insensitive to such a repetition. For false-negative case 3, the word *and* may have been commonly misreported as *an* in our raw training dataset. Overall, errored sentences in the ground truth may have reduced the model's ability to recognize true errors.

Our study had several limitations. First, only one dictation software (PowerScribe; Nuance Communications) and two medical institutions were included in this study, so the distribution of error types and consequently model performance may vary at other institutions with different software. However, Nuance is the predominant radiology SR software, holding 79% market share in 2018 (1), so the presented results are relevant to the majority of radiology workflows. Furthermore, the reports used represented the work of at least 152 unique radiologists. Second, the model used in this study is designed to flag one word in an incorrect phrase instead of the entire phrase, which slightly decreased the model's numerical performance on the prospective clinical test set. However, the main clinical purpose of the model is to bring potential errors to the attention of the radiologist, so flagging one word in an errored phrase fulfills this purpose. Third, using unscreened and therefore imperfect dictated radiology reports in the training set may have caused the model to learn to ignore some errors, leading to false-negative findings as discussed above. Fourth, BERT's pretraining framework did not allow the model to consider the context of the patient's electronic medical record, full report outside of the impression section, prior reports, or associated imaging when analyzing a sentence for errors. As discussed above, this technical limitation led to underestimation of model performance on the prospectively collected dataset. Experimenting with technical approaches, including RoBERTa (22), XLNet (23), and ALBERT (24), and additional data modalities (eg, imaging, other electronic health record text) could be goals for future studies, although currently limited by availability of data, high computational cost, and potentially inconsistent electronic health record information.

In conclusion, we have developed and evaluated a radiology domain-specific bidirectional transformer approach that could be used to detect and potentially correct SR errors. Other future work includes developing a more comprehensive error generator to improve the quality of training data and validating performance on multiple SR software and clinical workflows. As NLP methods continue to advance in their ability to extract contextual information, they can further reduce the proofreading burden of radiologists and improve the quality of radiology reports.

Acknowledgment: We thank Ashwin Balasubramaniam, MD, for his help with evaluation data collection.

Author contributions: Guarantors of integrity of entire study, **G.R.C., T.L.C., J.H.S.**; 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, **G.R.C., T.L., T.L.C., G.B.J., Y.S., J.H.S.**; clinical studies, **G.R.C., Y.J.L., T.H.V., J.H.S.**; experimental studies, **G.R.C., T.L., T.L.C., T.H.V., A.M.R., J.H.S.**; statistical analysis, **G.R.C., T.L., T.L.C., G.B.J., C.E.M., J.H.S.**; and manuscript editing, **G.R.C., T.L., T.L.C., G.B.J., M.V., T.H.V., Y.S., A.M.R., C.E.M., J.H.S.**

Disclosures of conflicts of interest: **G.R.C.** No relevant relationships. **T.L.** No relevant relationships. **T.L.C.** No relevant relationships. **G.B.J.** NIH funding. **M.V.** Consultant for USMLE Rx (acted as a consultant for the creation of medical student educational materials from 10/2020-5/2021. Payments made to author personally). **Y.J.L.** No relevant relationships. **T.H.V.** No relevant relationships. **Y.S.** No relevant relationships. **A.M.R.** Carestream Health/RSNA Research Scholar Grant; former *Radiology: Artificial Intelligence* trainee editorial board member. **C.E.M.** NIH grant to institution. **J.H.S.** No relevant relationships.

References

1. Bikman J. Speech Recognition In Radiology. Reaction Data, Inc. Published 2018. Accessed June 17, 2021.
2. Prevedello LM, Ledbetter S, Farkas C, Khorasani R. Implementation of speech recognition in a community-based radiology practice: effect on report turnaround times. J Am Coll Radiol 2014;11(4):402–406.
3. Ringler MD, Goss BC, Bartholmai BJ. Syntactic and semantic errors in radiology reports associated with speech recognition software. Health Informatics J 2017;23(1):3–13.
4. Hammana I, Lepanto L, Poder T, Bellemare C, Ly MS. Speech recognition in the radiology department: a systematic review. Health Inf Manag 2015;44(2):4–10.
5. Gutierrez F, Dou D, de Silva N, Fickas S. Online Reasoning for Semantic Error Detection in Text. J Data Semant 2017;6(3):139–153.
6. Voll K, Atkins S, Forster B. Improving the utility of speech recognition through error detection. J Digit Imaging 2008;21(4):371–377.
7. Minn MJ, Zandieh AR, Filice RW. Improving Radiology Report Quality by Rapidly Notifying Radiologist of Report Errors. J Digit Imaging 2015;28(4):492–498.
8. Zech J, Forde J, Titano JJ, Kaji D, Costa A, Oermann EK. Detecting insertion, substitution, and deletion errors in radiology reports using neural sequence-to-sequence models. Ann Transl Med 2019;7(11):233.
9. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805. http://arxiv.org/abs/1810.04805. Posted October 11, 2018. Accessed June 23, 2020.
10. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234–1240.
11. Beltagy I, Lo K, Cohan A. SciBERT: A Pretrained Language Model for Scientific Text. arXiv preprint arXiv:1903.10676. http://arxiv.org/abs/1903.10676. Posted March 26, 2019. Accessed December 29, 2020.
12. Meng X, Ganoe CH, Sieberg RT, Cheung YY, Hassanpour S. Self-Supervised Contextual Language Representation of Radiology Reports to Improve the Identification of Communication Urgency. AMIA Jt Summits Transl Sci Proc 2020;2020:413–421.
13. Chen L, Shah R, Link T, Bucknor M, Majumdar S, Pedoia V. Bert model fine-tuning for text classification in knee OA radiology reports. Osteoarthritis Cartilage 2020;28(Supplement 1):S315–S316.
14. Datta S, Ulinski M, Godfrey-Stovall J, Khanpara S, Riascos-Castaneda RF, Roberts K. Rad-SpatialNet: A Frame-based Resource for Fine-Grained Spatial Relations in Radiology Reports. LREC Int Conf Lang Resour Eval Proc Int Conf Lang Resour Eval. NIH Public Access, 2020; 2251.
15. Datta S, Roberts K. A Hybrid Deep Learning Approach for Spatial Trigger Extraction from Radiology Reports. Proc Conf Empir Methods Nat Lang Process Conf Empir Methods Nat Lang Process. NIH Public Access, 2020; 50.
16. Bressem KK, Adams LC, Gaudin RA, et al. Highly accurate classification of chest radiographic reports using a deep learning natural language model pretrained on 3.8 million text reports. Bioinformatics 2021;36(21):5255–5261.
17. Ong CJ, Orfanoudaki A, Zhang R, et al. Machine learning and natural language processing methods to identify ischemic stroke, acuity and location from radiology reports. PLoS One 2020;15(6):e0234908.
18. Zhou L, Blackley SV, Kowalski L, et al. Analysis of Errors in Dictated Clinical Documents Assisted by Speech Recognition Software and Professional Transcriptionists. JAMA Netw Open 2018;1(3):e180530.
19. Alsentzer E, Murphy JR, Boag W, et al. Publicly Available Clinical BERT Embeddings. arXiv preprint arXiv:1904.03323. http://arxiv.org/abs/1904.03323. Posted April 6, 2019. Accessed July 14, 2020.
20. Wolf T, Debut L, Sanh V, et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv preprint arXiv:1910.03771. http://arxiv.org/abs/1910.03771. Posted October 9, 2019. Accessed December 20, 2020.
21. Du Z, Qian Y, Liu X, et al. All NLP Tasks Are Generation Tasks: A General Pretraining Framework. arXiv preprint arXiv:2103.10360. http://arxiv.org/abs/2103.10360. Posted March 18, 2021. Accessed April 27, 2021.
22. Liu Y, Ott M, Goyal N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692. http://arxiv.org/abs/1907.11692. Posted July 26, 2019. Accessed June 23, 2020.
23. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV. XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv preprint arXiv:1906.08237. http://arxiv.org/abs/1906.08237. Posted June 19, 2019. Accessed December 29, 2020.
24. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. arXiv preprint arXiv:1909.11942. http://arxiv.org/abs/1909.11942. Posted September 26, 2019. Accessed December 29, 2020.