Radiology: Artificial Intelligence

TECHNICAL DEVELOPMENT

Automatic Diagnosis Labeling of Cardiovascular MRI by Using Semisupervised Natural Language Processing of Text Reports

*Sameer Zaman, MBBS, MRCP** • *Camille Petri, MD, MSc** • *Kavitha Vimalesvaran, MBBS, MRCP* • *James Howard, MB BChir, PhD* • *Anil Bharath, PhD* • *Darrel Francis, MB BChir, MD, FRCP* • *Nicholas S. Peters, MBBS, MD, FRCP* • *Graham D. Cole, MB BChir, PhD* • *Nick Linton, MBBS, PhD*

From the National Heart and Lung Institute, Imperial College London, Hammersmith Hospital, Du Cane Road, Second Floor B Block, London W12 0HS, England (S.Z., C.P., K.V., J.H., D.F., N.S.P., G.D.C.); Imperial College Healthcare National Health Service Trust, London, England (J.H., D.F., N.S.P., G.D.C., N.L.); and Department of Bioengineering, Imperial College London, London, England (A.B., N.L.). Received March 23, 2021; revision requested May 3; revision received October 29; accepted November 3. **Address correspondence to** G.D.C. (e-mail: *g.cole@imperial.ac.uk*).

S.Z. and C.P. supported by the UK Research and Innovation Centre for Doctoral Training in Artificial Intelligence for Healthcare (grant EP/S023283/1).

*S.Z. and C.P. contributed equally to this work.

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

Radiology: Artificial Intelligence 2022; 4(1):e210085 • https://doi.org/10.1148/ryai.210085 • Content codes: AI CA MR

Purpose: To assess whether the semisupervised natural language processing (NLP) of text from clinical radiology reports could provide useful automated diagnosis categorization for ground truth labeling to overcome manual labeling bottlenecks in the machine learning pipeline.

Materials and Methods: In this retrospective study, 1503 text cardiac MRI reports from 2016 to 2019 were manually annotated for five diagnoses by clinicians: normal, dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy, myocardial infarction (MI), and myocarditis. A semisupervised method that uses bidirectional encoder representations from transformers (BERT) pretrained on 1.14 million scientific publications was fine-tuned by using the manually extracted labels, with a report dataset split into groups of 801 for training, 302 for validation, and 400 for testing. The model's performance was compared with two traditional NLP models: a rule-based model and a support vector machine (SVM) model. The models' F1 scores and receiver operating characteristic curves were used to analyze performance.

Results: After 15 epochs, the F1 scores on the test set of 400 reports were as follows: normal, 84%; DCM, 79%; hypertrophic cardiomyopathy, 86%; MI, 91%; and myocarditis, 86%. The pooled F1 score and area under the receiver operating curve were 86% and 0.96, respectively. On the same test set, the BERT model had a higher performance than the rule-based model (F1 score, 42%) and SVM model (F1 score, 82%). Diagnosis categories classified by using the BERT model performed the labeling of 1000 MR images in 0.2 second.

Conclusion: The developed model used labels extracted from radiology reports to provide automated diagnosis categorization of MR images with a high level of performance.

Supplemental material is available for this article.

An earlier incorrect version of this article appeared online. This article was corrected on April 12, 2022.

A major challenge for researchers building machine learning algorithms for medical imaging is the great expense of ground truth labeling, even if labeling consists of only manual classification of clinical reports that are already written. If the process of creating ground truth labels from clinical radiology reports could be automated, the bottleneck of extensive labeling within the clinical research pipeline could be relieved (1). Additionally, automatic labeling would also capitalize on the many skilled clinician hours that were spent generating clinical reports.

Although deep learning techniques have been applied to cardiac MRI (CMR), prior studies have largely focused on automating measurements from images that are documented in clinical reports in a standardized format, which can easily be extracted as ground truth labels (2,3). However, the most important clinical questions from a CMR acquisition are not measurements but are rather diagnoses, such as the presence of scars that distinguish between different cardiac conditions (4–6). Such questions are more difficult for machine learning researchers to address because of the need to manually review clinical reports to convert them into diagnostic categories suitable for use as the ground truth.

Natural language processing (NLP) can automate the process of converting text into diagnoses. Deep learning techniques perform well (7) but require datasets with many manually annotated data points (1), which can prevent the widespread application of these models in clinical research. Bidirectional encoder representations from transformers (BERT) models (8), which use transfer learning, can be pretrained on a large unlabeled text corpus. Users can then take a version "off the shelf" and fine-tune the model for a specific task. This semisupervised learning requires much less manual labeling, without compromising algorithmic performance (1,9).

Abbreviations

AUC = area under the receiver operating characteristic curve, BERT = bidirectional encoder representations from transformers, CMR = cardiac MRI, DCM = dilated cardiomyopathy, MI = myocardial infarction, NLP = natural language processing, SVM = support vector machine

Summary

A semisupervised natural language processing (NLP) algorithm, based on bidirectional transformers, accurately categorized diagnoses from cardiac MRI text of radiology reports for the labeling of MR images; the model had a higher accuracy than traditional NLP models and performed faster labeling than clinicians.

Key Points

- A bidirectional transformer–based natural language processing model performed well at automatically categorizing five diagnoses from cardiac MRI text reports (F1 score, 86%; pooled area under the receiver operating characteristic curve, 0.96).
- The pretrained model required only approximately 800 reports, which were manually labeled by clinicians, for fine-tuning.
- The model annotated 1000 MRI acquisitions with diagnosis labels in 0.2 second.

Keywords

Semisupervised Learning, Diagnosis/Classification/Application Domain, Named Entity Recognition, MRI

In this study, we applied a BERT-based model and compared it with two well-established NLP approaches that share a similar programming time frame and computational power demand: a rule-based approach and a support vector machine (SVM). We demonstrate a research application of BERT that can perform rapid diagnosis classification from radiology reports, which can be used as the ground truth for imaging machine learning research.

Materials and Methods

Ethical Approval

Ethical approval was gained from the UK Health Regulatory Agency (Integrated Research Application System identifier 243023), and consent was waived.

Data Extraction and Preprocessing

In this retrospective study, text CMR reports produced between 2016 and 2019 at three hospitals in London, England, were retrospectively extracted as plain text (.txt) files from the electronic health record. Reports were reviewed and annotated by clinicians (S.Z., K.V.), with at least 3 years of experience reviewing CMR reports, by using an internal glossary and labeling criteria established at a consensus meeting of authors. The rationale for diagnosis labeling is described in Appendix E1 (supplement). Five diagnoses were manually labeled: normal, dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy, myocardial infarction (MI), and myocarditis. Diagnoses were labeled if, in the report, the clinician indicated that it was a likely explanation for the image findings according to international clinical guidelines and diagnostic criteria. Discrepancies were addressed at a meeting between the authors to achieve consensus. The normal classification meant there was no clinically relevant cardiac abnormality in the text report (Appendix E1 [supplement]). A total of 1600 reports were initially screened for inclusion, and 1503 were ultimately included after excluding duplicate and incomplete reports.

Intra- and Interrater Variability

A sample of 100 reports selected by stratified random sampling was double labeled by an experienced clinician (S.Z.) at least 2 weeks apart in a blinded manner. The same 100 reports were labeled by a second blinded clinician (K.V.). Intra- and interrater variability were assessed by calculating Cohen κ coefficients (Appendix E2 [supplement]).

BERT Model Machine Learning Architecture

A total of 1503 labeled CMR reports were included for model development and were split into training (n = 801), validation (n = 302), and testing (n = 400) datasets (Fig 1). The approach consisted of multilabel document-level classification tasks, with each class being a diagnosis of interest. Each report could be classified in multiple classes. Preprocessing included anonymization, practitioner information removal, nonalphanumeric character removal, decapitalization, and white-space removal.

The model architecture was based on SciBERT (10), a version of BERT (8) pretrained for a language-modeling task on a corpus of 1.14 million scientific publications. BERT models rely on a multihead attention architecture, which enables them to manage multiple dependencies of different lengths in the input sequence (8).

Our documents were tokenized by using a byte pair encodinglike approach (11). A *token* is a group of characters packaged together as one unit for processing (Appendix E3 [supplement]). The tokenized versions of the documents were used as inputs for the SciBERT model. Each token corresponds to a unique token identifier, which is a natural number. Those token identifiers are used as inputs, with an upper limit of 512 input tokens for the model. In a document dataset of size $N \in \mathbb{N}$, for a tokenized document $d_i \in \mathbb{N}^m$, $i \in \{1,2,...,N\}$ consisting of $m \in \{1,2,...,512\}$ token identifiers, SciBERT generates r hidden representations $H_t \in \mathbb{R}^{m \times 768}$, $t \in \{1,2,...,r\}$ that can be pooled into a vector $v_{d_i} \in \mathbb{R}^{768}$ of constant dimension for every document. This vector v_{d_i} was then used as the input for a classifier by using a fully connected layer with a sigmoid activation function $\sigma(.)$ from which the output was a vector of likelihood for each diagnosis $\grave{y}_{d_i} \in [0,1]^c$, with c being the number of classes, following the formula:

$$\grave{y}_{d_i} = \sigma(A^T v_{d_i} + b),$$

where $A \in \mathbb{R}^{768 \times c}$ and $b \in \mathbb{R}^c$.

The diagnosis classes within our dataset were heavily imbalanced. To account for class imbalance during training, we defined a weighted binary cross-entropy loss function (Appendix E2 [supplement]). During fine-tuning for the classification task using our training dataset, all the trainable parameters from SciBERT and the classifier were updated. During training, the performance was evaluated on the validation set at each epoch by

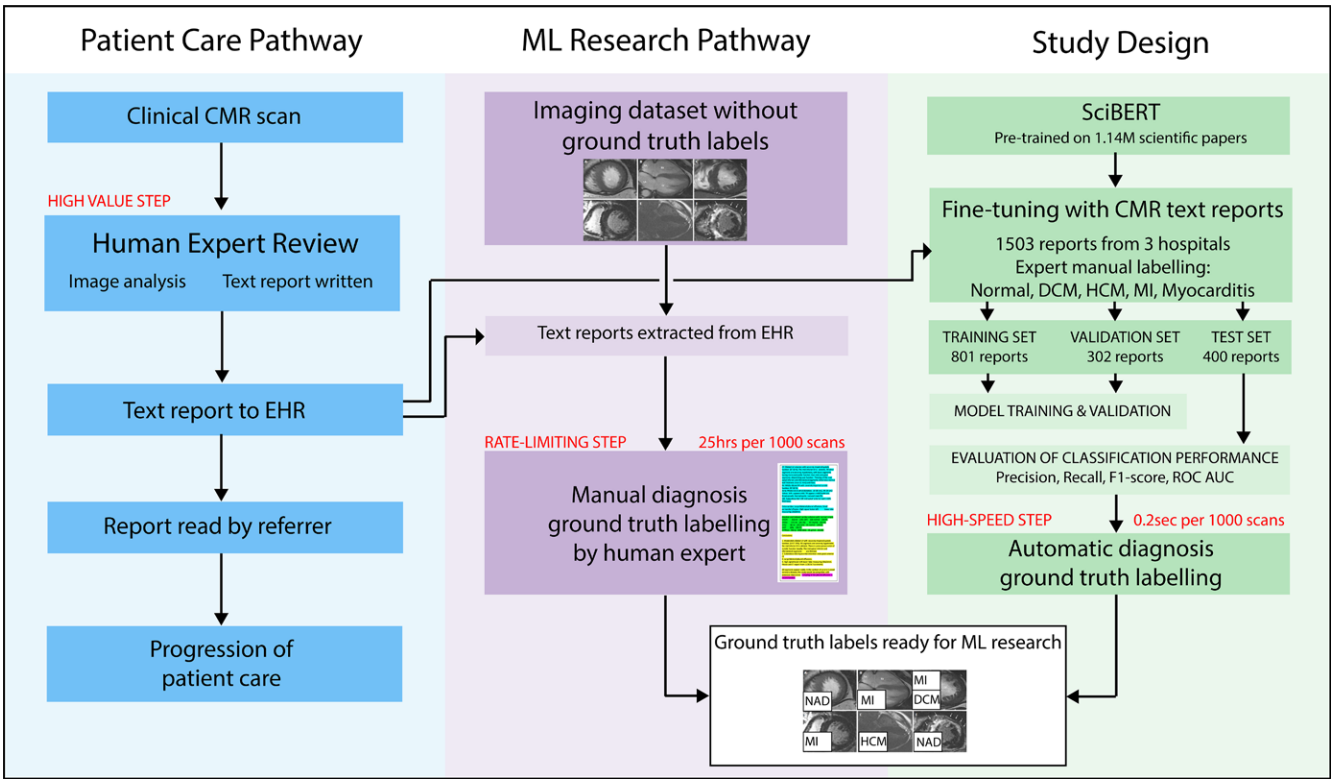

Figure 1: Data pipeline of clinical CMR acquisitions in routine clinical practice and for ML research, along with study design. AUC = area under the ROC curve, BERT = bidirectional encoder representations from transformers, CMR = cardiac MRI, DCM = dilated cardiomyopathy, EHR = electronic health record, HCM = hypertrophic cardiomyopathy, MI = myocardial infarction, ML = machine learning, ROC = receiver operating characteristic.

using a microaveraged F1 score (the harmonic mean of recall and precision). Microaveraging (rather than macroaveraging) and F1 scores (rather than overall accuracy) were chosen because these metrics better account for diagnosis class imbalances (12). The model with the highest F1 score was saved for further evaluation.

A minibatch gradient descent approach was used for optimization. The model with the highest performance was obtained by using the Adam with weight decay optimizer (13) (Appendix E2 [supplement]). Gradient clipping with a maximum gradient norm of 1 was used to prevent model saturation. The model was developed by using the PyTorch framework (14), and the SciBERT model was implemented in the Hugging Face framework (15). Performance was evaluated on the test set with a decision threshold of 0.5 for each prediction. The model was developed with a NVIDIA P1000 graphics processing unit by using bespoke Python 3 scripts (16).

First 512 versus Last 512 Tokens

BERT has a maximum length limit of 512 tokens (8), which is often exceeded by radiology reports. We trained two versions of the model, taking either the first 512 tokens or the last 512 tokens within each report as model inputs. We compared the performance of both model versions on the test set.

Baseline NLP Models for Benchmarking

Rule-based NLP and SVM models were developed to benchmark the performance of the BERT model. These models were chosen because they are the most realistic alternatives to BERT for clinical researchers performing ground truth labeling from text with relatively little programming time and computational expense. Both models were designed as discrete binary classifiers that did not put out a probability.

Rule-based NLP model.— All possible expressions used to identify each diagnosis were established at a consensus meeting of cardiologists (S.Z., K.V., D.F., G.D.C., N.L.). The decision rule of the model was programmed to classify a report as positive for a particular diagnosis label if any instance of those specific terms or expressions appeared anywhere in the report text.

SVM model.— An SVM model was trained for a one-versus-all classification task by using high-level features as its input. The input text was tokenized and lemmatized, and a term frequency (inverse document frequency) measure was used for text vectorization. An SVM for a classification submodel was trained independently for each label in a one-versus-all fashion. The training set was used to optimize the parameters of the separating hyperplane that showed the highest performance on the validation set for each label. The best-performing model used a Gaussian kernel, accounting for nonlinearity and class imbalances.

Statistical Analysis

Because of class imbalances in the dataset, F1 scores were used as the primary performance metric. For the BERT-based model, the area under the receiver operating charac-

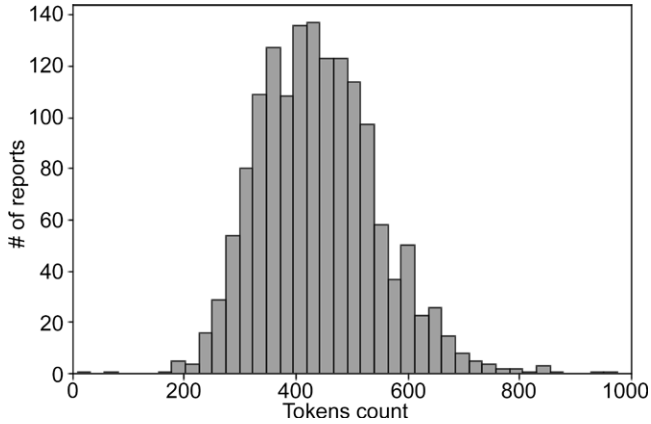

Figure 2: Distribution of the number of tokens (words) per report in the dataset of 1503 cardiac MRI text reports.

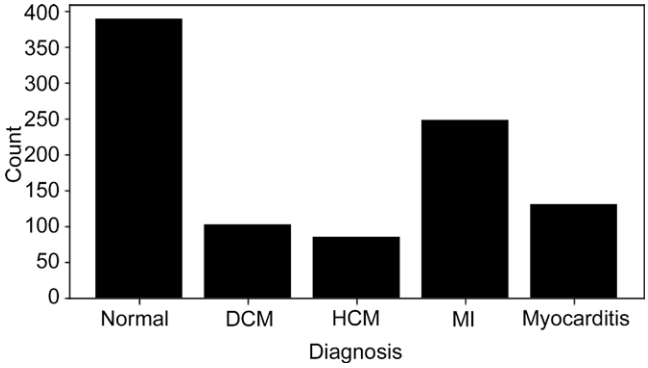

Figure 3: Frequency of the five diagnosis categories in the dataset of 1503 free-text cardiac MRI reports. DCM = dilated cardiomyopathy, HCM = hypertrophic cardiomyopathy, MI = myocardial infarction.

Table 1: Precision, Recall, and F1 Scores for the BERT-based, NLP Rule-based, and SVM Models on the Test Set

Parameter	BERT (%)	Rule-based (%)	SVM (%)
Precision			
Normal (*n* = 110)	87	28	76
DCM (*n* = 27)	81	85	77
HCM (*n* = 22)	90	73	90
MI (*n* = 68)	97	17	91
Myocarditis (*n* = 29)	89	41	65
Microaveraged	89	27	79
Recall			
Normal (*n* = 110)	81	100	89
DCM (*n* = 27)	78	85	85
HCM (*n* = 22)	82	86	82
MI (*n* = 68)	85	100	85
Myocarditis (*n* = 29)	83	83	76
Microaveraged	82	95	86
F1 score			
Normal (*n* = 110)	84	43	82
DCM (*n* = 27)	79	85	81
HCM (*n* = 22)	86	79	86
MI (*n* = 68)	91	30	88
Myocarditis (*n* = 29)	86	55	70
Microaveraged	86	42	82

Note.—Models were assessed on the test dataset of 400 reports. BERT = bidirectional encoder representations from transformers, DCM = dilated cardiomyopathy, HCM = hypertrophic cardiomyopathy, MI = myocardial infarction, SVM = support vector machine.

teristic curve (AUC) and the 95% CI were also computed. The baseline models were discrete classifiers (did not output a probability), so their performance was not represented by a receiver operating characteristic curve. AUCs were derived for the baseline models by making assumptions about the shape of the hidden curve, as described by van den Hout (17). A pairwise comparison of each baseline AUC to the BERT-based model was made by using DeLong tests. A *P* value $< .05$ was considered to indicate a significant difference. Results for these comparisons are provided in Appendix E4 (supplement).

Model Availability

The code for these methods and other resources is shared publicly in an online repository (*https://github.com/cpetri/CMR_NLP_BERT*).

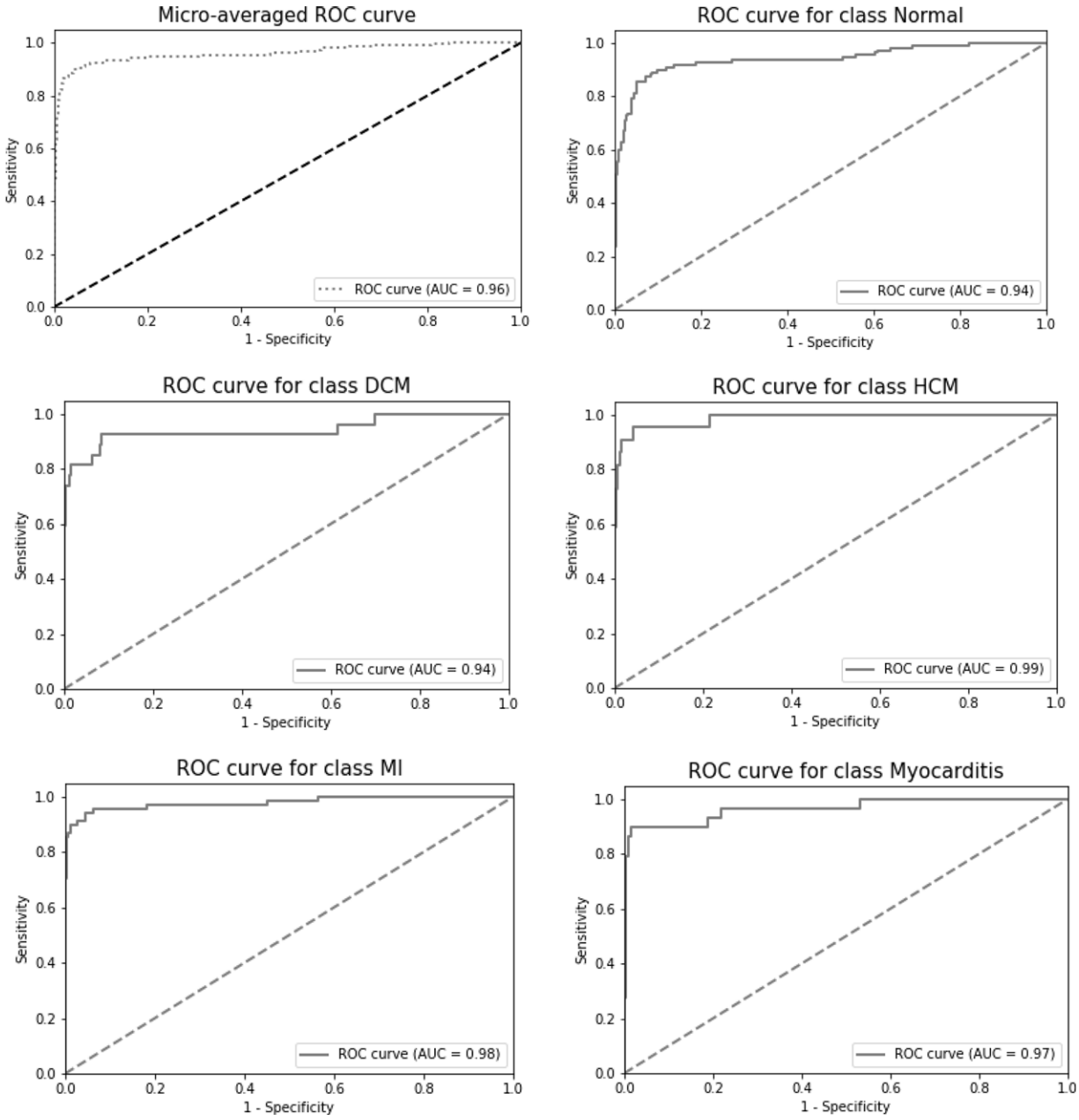

Figure 4: ROCs of the bidirectional encoder representations from transformers–based model, microaveraged across all diagnosis classes (top left panel) and for each individual diagnosis class. The AUC is displayed in the legends. AUC = area under the ROC curve, DCM = dilated cardiomyopathy, HCM = hypertrophic cardiomyopathy, MI = myocardial infarction, ROC = receiver operating characteristic.

Results

Performance of the BERT-based Model and Baseline Models

The distribution of the tokenized length of documents is shown in Figure 2. Diagnosis class frequencies in the entire dataset are shown in Figure 3. Because of class imbalances in the data, microaveraging and F1 scores were used to report results rather than macroaveraging and overall accuracy, which would mainly reflect accuracy in the dominant class. The BERT-based model achieved a microaveraged F1 score of 86%. This was higher than F1 scores for the rule-based NLP (42%) and SVM (82%) models. The precision, recall, and F1 scores for each diagnosis are shown in Table 1.

For the BERT-based model, the AUC was computed for each diagnosis by using a one-versus-all approach (Fig 4). The BERT-based model had an AUC of at least 0.94 for every diagnosis and a microaveraged AUC of 0.96 (95% CI: 0.94, 0.98) (Table 2). Because our rule-based NLP and SVM models were discrete classifiers that did not output a probability, the performance of these models is represented by one point rather than a curve (Appendix E4, Fig E2 [supplement]); however, an AUC can be derived by using previously described methods (17). The BERT-based model had a higher AUC than AUCs derived for the rule-based NLP (AUC, 0.79 [95% CI: 0.77, 0.80]; $P < .01$) and SVM (AUC, 0.91 [95% CI: 0.89, 0.93]; $P < .01$) models (Appendix E4, Table E3 [supplement]). These

results should be cautiously interpreted because of the assumptions about hidden receiver operating characteristic curves required to calculate the AUC for the baseline models.

Inter- and Intrarater Variability

Intrarater variability in 100 reports labeled twice by the same clinician was low (pooled Cohen κ = 0.95) (Table 3). Interrater variability for the same 100 reports was also low (pooled Cohen κ = 0.93) (Table 3).

Table 2: AUCs of the BERT-based Model on the Test Set of 400 Reports

Diagnosis	AUC
Normal (*n* = 110)	0.94 (0.91,0.97)
DCM (*n* = 27)	0.94 (0.87, 0.99)
HCM (*n* = 22)	0.99 (0.96, 1.00)
MI (*n* = 68)	0.98 (0.95, 1.00)
Myocarditis (*n* = 29)	0.97 (0.92, 1.00)
Microaveraged	0.96 (0.94, 0.98)

Note.—Data in parentheses in the AUC column are 95% CIs. AUC = area under the receiver operating characteristic curve, BERT = bidirectional encoder representations from transformers, DCM = dilated cardiomyopathy, HCM = hypertrophic cardiomyopathy, MI = myocardial infarction.

First 512 versus Last 512 Tokens

Overall, the first 512 version (F1 score, 86%) had a higher performance than the last 512 version (F1 score, 81%). The first 512 model had a higher performance than the last 512 model for all diagnoses except MI (Table 4).

Discussion

We demonstrated that semisupervised NLP can be used for automatically extracting ground truth diagnosis labels from free-text CMR reports with relatively little clinician effort and computational expense. Our BERT-based model demonstrated good performance compared with clinician labeling and two alternative NLP models used to classify five cardiology diagnoses (F1 score, 86%; AUC, 0.96).

Our BERT-based model assigned diagnosis labels to 1000 MR images in 0.2 second. The implementation of this model within the clinical machine learning pipeline would help alleviate manual labeling bottlenecks. The data requirements for deep learning models are usually overwhelmingly high for clinicians to label within a reasonable resource expense (18). The BERT model overcomes this time expense with transfer learning, which includes pretraining on a large corpus of scientific language, and then fine-tuning for the specific task, which requires substantially less manual annotation (20 hours for our training set). We propose that this model could be repurposed for other

Table 3: Intra- and Interrater Variability of Diagnosis Labels in a Sample of 100 Text Reports

Diagnosis	Clinician 1 (Ground Truth)	Clinician 1 (Blinded)	Clinician 2 (Blinded)	Intrarater Variability	Interrater Variability
Normal	31	30	32	0.93	0.88
DCM	7	7	7	1.00	1.00
HCM	6	4	7	0.79	0.92
MI	15	15	14	1.00	0.88
Myocarditis	10	10	11	1.00	0.95
Average	…	…	…	0.95	0.93

Note.—The number of reports for each diagnosis for each clinician are shown. Cohen κ statistic was used to assess intra- and interrater variability. DCM = dilated cardiomyopathy, HCM = hypertrophic cardiomyopathy, MI = myocardial infarction.

Table 4: Comparison of the F1 Scores of Two Versions of the BERT-based Model

Diagnosis (BERT Model)	First 512 Tokens, F1 Score (%)	Last 512 Tokens, F1 Score (%)
Normal (*n* = 110)	84	80
DCM (*n* = 27)	79	56
HCM (*n* = 22)	86	70
MI (*n* = 68)	91	95
Myocarditis (*n* = 29)	86	79
Microaveraged	86	81

Note.—The two models described here used a different set of tokens as model inputs (ie, the first 512 and the last 512 tokens). BERT = bidirectional encoder representations from transformers, DCM = dilated cardiomyopathy, HCM = hypertrophic cardiomyopathy, MI = myocardial infarction.

modalities, given that the model requires relatively few labeled data points for training.

Rule-based and deep learning NLP models have been used for clinical text categorization (19–22), but both can suffer a decrease in performance when a model developed in one clinical area is applied to a different area (23). It is not known whether the same performance drop would occur in trained BERT models, which are deeply bidirectional. The bidirectionality of the BERT model means that the model uses context both before and after words to associate a specific meaning (ie, labels) to the words, thus learning relationships within sentences. In addition, the BERT model also learns relationships between sentences and thus is trained to have a highly interwoven network of connected and related words (24).

Clinical researchers considering quick and inexpensive ways to automate ground truth labeling from text may use alternative rule-based approaches or SVMs. Overall, the BERT model had a higher performance than both the rule-based NLP model and the SVM model in the primary performance measure of F1 scores. The rule-based model had a high sensitivity (average of 95%) because it was designed to assign a label whenever it found any one of the predefined ways of saying the diagnosis (eg, *infarction*, *myocardial infarction*, *infarct*, *MI*). This method is akin to search functions or IF commands, commonly used by researchers searching data for diagnoses. This method incorrectly labeled instances of negation (eg, there is no myocardial infarction), causing poor overall specificity (27%). This method would label many false-positive results and would require manual efforts to remove. The rule-based NLP model could be fine-tuned to improve its specificity (eg, excluding diagnoses preceded by words of negation), but this modification might, in turn, create a model highly biased to specific reporting styles.

The SVM model showed F1 scores comparable with those of the BERT model for all diagnoses except myocarditis. This result may have been due to the fact that the CMR features of myocarditis overlap with those of other conditions, making it a more ambiguous diagnosis in clinical practice. The BERT model may better account for linguistic uncertainty by being deeply bidirectional. Statistical analyses showed that the BERT model had significantly higher AUCs than the SVM model for all diagnoses except DCM (Appendix E4, Table E3 [supplement]). These results should be interpreted more cautiously because of the assumptions made to calculate the AUC of the baseline model. However, this trend was also reflected in the F1 scores for DCM (BERT, 79%; rule-based NLP, 85%; and SVM, 81%), which might be because, in clinical practice, DCM is rarely mentioned in a report, other than when stating that it is present, so traditional NLP can detect it relatively easily with a low false-positive rate.

A total of 466 of 1503 (31%) of the reports exceeded 512 words. The first 512 version of the BERT model had higher F1 scores than the last 512 version for all diagnoses except MI. This result is surprising because the final part of the report contains the "conclusions" section, in which diagnoses are listed. MI has a highly biased CMR appearance, so reporters may more explicitly state this diagnosis in their conclusions. Other diagnoses can be more ambiguous, perhaps explaining why the BERT model had a lower performance when only the last 512 words of the reports were used. A longer report may also reflect more uncertainty in the case overall; thus, reports with longer conclusions may be less critical for the model than shorter ones.

Our study had limitations. Our BERT model was not 100% accurate for all five diagnoses. Some diagnostic uncertainty routinely occurs in clinical CMR (25) (interrater Cohen $\kappa = 0.93$). Achieving consensus for research labeling is a laborious and time-consuming process. The BERT model agreed with clinicians 92% as often as the clinicians agreed with each other. In addition, the model produced diagnoses faster than clinicians. Our model performed well at classifying diagnoses from text CMR reports. Whether this performance is generalizable to other modalities with minimal fine-tuning is not known. We have made our code publicly available so other researchers can use our model. We did not explore alternative ways to overcome the 512-token input limitation of BERT; however, these are proposed elsewhere (26).

Our BERT-based, semisupervised NLP algorithm can accurately perform automatic diagnosis categorization of text CMR reports. Being able to use this model to automate the ground truth diagnosis labeling of MR images would substantially reduce clinician efforts in machine learning research.

Author contributions: Guarantors of integrity of entire study, **S.Z., C.P., G.D.C.**; 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; literature research, **S.Z., C.P., G.D.C., N.L.**; clinical studies, **S.Z., K.V., J.H., N.L.**; experimental studies, **S.Z., K.V.**; statistical analysis, **S.Z., C.P., J.H., G.D.C., N.L.**; and manuscript editing, all authors

Disclosures of conflicts of interest: **S.Z.** Funding paid to institution (Imperial College London) by UK Research and Innovation (grant no EP/S023283/1). **C.P.** Funding paid to institution (Imperial College London) by UK Research and Innovation (grant no EP/S023283/1). **K.V.** No relevant relationships. **J.H.** Institution receives Wellcome Trust Research Grant for author's salary; support for attending meetings and/or travel from the Society for Cardiovascular Magnetic Resonance Travel Scholarship; *Radiology: Artificial Intelligence* trainee editorial board member. **A.B.** Institution supported by Rosetrees Trust, UKRI, DAMAE Medical, Imperial College, and Wellcome Trust; nonremunerated roles as part of normal engineering academic position with Fight for Sight, Data for Policy, and Association of City and Guilds. **D.F.** No relevant relationships. **N.S.P.** No relevant relationships. **G.D.C.** No relevant relationships. **N.L.** Work supported by funding from AI4Health (funding of PhD fellowship).

References

1. 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.
2. Howard JP, Zaman S, Ragavan A, et al. Automated analysis and detection of abnormalities in transaxial anatomical cardiovascular magnetic resonance images: a proof of concept study with potential to optimize image acquisition. Int J Cardiovasc Imaging 2021;37(3):1033–1042.
3. Martini N, Aimo A, Barison A, et al. Deep learning to diagnose cardiac amyloidosis from cardiovascular magnetic resonance. J Cardiovasc Magn Reson 2020;22(1):84.
4. Weng Z, Yao J, Chan RH, et al. Prognostic value of LGE-CMR in HCM: a meta-analysis. JACC Cardiovasc Imaging 2016;9(12):1392–1402.
5. Becker MAJ, Cornel JH, van de Ven PM, van Rossum AC, Allaart CP, Germans T. The prognostic value of late gadolinium-enhanced cardiac magnetic resonance imaging in nonischemic dilated cardiomyopathy: a review and meta-analysis. JACC Cardiovasc Imaging 2018;11(9):1274–1284.
6. Raina S, Lensing SY, Nairooz RS, et al. Prognostic value of late gadolinium enhancement CMR in systemic amyloidosis. JACC Cardiovasc Imaging 2016;9(11):1267–1277.

7. Sorin V, Barash Y, Konen E, Klang E. Deep learning for natural language processing in radiology—fundamentals and a systematic review. J Am Coll Radiol 2020;17(5):639–648.
8. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. ArXiv 1810.04805 [preprint] http://arxiv.org/abs/1810.04805. Posted May 24, 2019. Accessed July 26, 2020.
9. Liu H, Zhang Z, Xu Y, et al. Use of BERT (bidirectional encoder representations from transformers)-based deep learning method for extracting evidences in Chinese radiology reports: development of a computer-aided liver cancer diagnosis framework. J Med Internet Res 2021;23(1):e19689.
10. Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. ArXiv 1903.10676 [preprint] https://arxiv.org/abs/1903.10676. Posted September 10, 2019. Accessed February 1, 2021.
11. Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. ArXiv 1508.07909 [preprint] https://arxiv.org/abs/1508.07909. Posted June 10, 2016. Accessed March 23, 2021.
12. Singh PK, Kar AK, Singh Y, Kolekar MH, Tanwar S, eds. Proceedings of ICRIC 2019: recent innovations in computing. Cham, Switzerland: Springer, 2020.
13. Loshchilov I, Hutter F. Decoupled weight decay regularization. ArXiv 1711.05101 [preprint] http://arxiv.org/abs/1711.05101. Posted January 4, 2019. Accessed March 23, 2021.
14. Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library. ArXiv 1912.01703 [preprint] http://arxiv.org/abs/1912.01703. Posted December 3, 2019. Accessed March 23, 2021.
15. Wolf T, Debut L, Sanh V, et al. HuggingFace's transformers: state-of-the-art natural language processing. ArXiv 1910.03771 [preprint] http://arxiv.org/abs/1910.03771. Posted July 13, 2020. Accessed March 23, 2021.
16. Van Rossum G, Drake FL. Python 3 reference manual. Scotts Valley, Calif: CreateSpace, 2009.
17. van den Hout WB. The area under an ROC curve with limited information. Med Decis Making 2003;23(2):160–166.
18. Wang Y, Sohn S, Liu S, et al. A clinical text classification paradigm using weak supervision and deep representation. BMC Med Inform Decis Mak 2019;19(1):1.
19. Klang E. Deep learning and medical imaging. J Thorac Dis 2018;10(3):1325–1328.
20. Carrodeguas E, Lacson R, Swanson W, Khorasani R. Use of machine learning to identify follow-up recommendations in radiology reports. J Am Coll Radiol 2019;16(3):336–343.
21. Lee C, Kim Y, Kim YS, Jang J. Automatic disease annotation from radiology reports using artificial intelligence implemented by a recurrent neural network. AJR Am J Roentgenol 2019;212(4):734–740.
22. Banerjee I, Ling Y, Chen MC, et al. Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification. Artif Intell Med 2019;97:79–88.
23. Ye Y, Wagner MM, Cooper GF, et al. A study of the transferability of influenza case detection systems between two large healthcare systems. PLoS One 2017;12(4):e0174970.
24. Alammar J. The Illustrated BERT, ELMo, and co. (how NLP cracked transfer learning). Visualizing machine learning one concept at a time. Jay Alammar. https://jalammar.github.io/illustrated-bert/. Published Dec 3, 2018. Updated 2021. Accessed July 26, 2020.
25. Juneau D, Nery PB, Pena E, et al. Reproducibility of cardiac magnetic resonance imaging in patients referred for the assessment of cardiac sarcoidosis; implications for clinical practice. Int J Cardiovasc Imaging 2020;36(11):2199–2207.
26. Fiok K, Karwowski W, Gutierrez E, et al. Text guide: improving the quality of long text classification by a text selection method based on feature importance. ArXiv 2104.07225 [preprint] http://arxiv.org/abs/2104.07225. Posted April 15, 2021. Accessed July 5, 2021.