Title: EchoGraph: A Specialized Solution for Automatic Echocardiography Report Quality Evaluation

Authors: Chieh-Ju Chao, MD[1,2], Jean-Benoit Delbrouck, PhD[3], Mohammand Asadi[4], Imon Banerjee, PhD[5], Juan M. Farina, MD[6], Francesca Galasso[6], Ahmed K Mahmoud, MD[6], Mohammed Tisser Abbas, MD[6], Reza Arsanjani, MD[6], Garvan C Kane, MD, PhD[1], Jae K Oh, MD[1], Bradley J Erickson, MD, PhD[7], Li Fei-Fei, PhD[2], Ehsan Adeli, PhD[2], and Curtis Langlotz, MD, PhD[3]

[1] Department of Cardiovascular Medicine, Mayo Clinic, Rochester, Minnesota, USA
[2] Institute of Human-centered Artificial Intelligence, Stanford University, Palo Alto, California, USA
[3] Center of Artificial Intelligence in Medical Imaging, Stanford University, Palo Alto, California, USA
[4] Department of Electrical Engineering, Stanford University, Palo Alto, California, USA
[5] Department of Radiology, Mayo Clinic, Scottsdale, Arizona, USA
[6] Department of Cardiovascular Diseases, Mayo Clinic, Scottsdale, Arizona, USA
[7] Department of Radiology, Mayo Clinic, Rochester, Minnesota, USA

Abbreviations:
AI: Artificial Intelligence
DL: Deep Learning
Echo: Echocardiography
FT: Fine-Tuning
GPT: Generative Pre-trained Transformer
NLP: Natural Language Processing
Seq2seq: Sequence-to-sequence
TTE: Transthoracic Echocardiography
TEE: Transesophageal Echocardiography
LLM: Large Language Model

Abstract
With the rise of generative AI, there is a growing need for automatic metrics to evaluate the factual accuracy of clinical text, yet no such tool exists for echocardiography. Existing evaluation tools often underperform in this domain. To address this, we developed EchoGraph, a BERT-based model trained on densely annotated echocardiography reports using a schema tailored for this subspecialty and a dedicated F1-style reward to emphasize clinically important components. EchoGraph demonstrated strong performance in predicting entities (micro F1 0.85) and relations (micro F1 0.70). Its F1 reward was more sensitive to detecting corrupted report content than the RadGraph F1 reward (showing a 50%-60% vs. 6%-12% drop). EchoGraph thus offers an effective solution for evaluating and advancing language model-based applications in echocardiography, supporting the development of more accurate and clinically meaningful AI-generated reports.

Introduction

Echocardiography (echo) reports are the primary communication tool for echo studies, containing detailed quantitative measurements and cardiologists' interpretations concerning each patient's medical history and clinical context[1,2]. Although these reports hold substantial clinically relevant information, their free-text format presents significant challenges for large-scale use. The unstructured nature of this data, combined with complex medical language, hampers automated extraction and analysis, limiting their utility in artificial intelligence (AI) research and clinical practice[3]. With the rise of generative AI, particularly language models[4,5], there is an increasing need for automatic, generic metrics to evaluate the factual correctness of generated clinical text to facilitate the pace of progress in developing vision-language models that produce accurate reports[3,6,7]. Studies have leveraged information extraction or language model techniques for the development of automatic metrics or frameworks to evaluate the reliability and accuracy of AI-generated clinical narratives, enhancing their utility in medical practice and research[3,6,8–10].

However, studies specifically for echo reports remain limited, and earlier works have not been developed using generic schemas for information extraction[11–13]. Specifically, these studies often focus on a relatively narrow component, such as the comparison statement of reports[13], or use pre-defined important elements that do not capture the full report content[11,12]. The scarcity of densely annotated reports likely led to reliance on narrow extraction schemas, limiting information scope and hindering the development of comprehensive metrics for assessing factual correctness in clinical echo reports.

RadGraph and RadGraph XL are knowledge graph models addressing similar challenges in radiology report analysis[3,6]. However, despite robust out-of-distribution performance, their direct application to echo reports has significant limitations[3,6,14–16]. Unlike chest X-rays, echo reports frequently contain detailed quantitative measurements that are critical for clinical decision-making, and RadGraph rewards lack the sensitivity to accurately detect and differentiate differences in quantitative measurements[16]. This highlighted the need for a dedicated model tailored specifically to the unique demands of echo reports.

In this context, we plan to develop EchoGraph with a refined clinical entity and relation extraction schema and reward function specifically tailored to echo reports. We hypothesize that the EchoGraph model, along with the new reward, will achieve comparable performance to RadGraph in detecting shared entities and relations while demonstrating enhanced sensitivity in accurately capturing quantitative measurements. This dedicated approach aims to address the unique requirements of unstructured echo report data, ensuring precise evaluation of clinically relevant metrics.

Methods

Dataset, Split, and Preprocessing

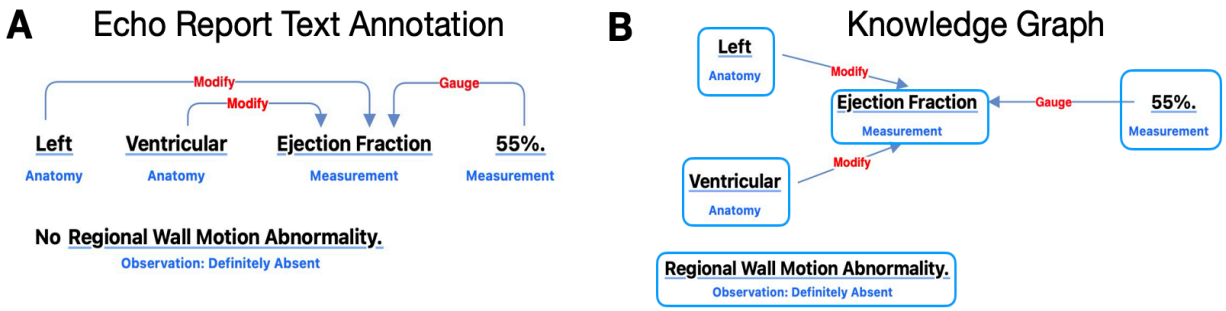
Mayo Clinic Echo Reports: Adult (> 18 years old) echo studies performed from 1/1/2017 to 12/31/2017 at the Mayo Clinic Enterprise were retrieved, and 600 reports from non-duplicated patients were randomly selected. The "Findings" section was used to ensure that all the relevant findings were covered. Subheadings were removed from the text. Following this, the dataset was split into the train, validation, and test sets in a 7:2:1 ratio. A validation set (n=1,000) from Mayo Arizona was randomly selected for the data corruption study; the characteristics of this set were described in our previous work[16].

MIMIC-III ECHO-NOTE2NUM Reports: (v.1.0.0, referred to as MIMIC-EchoNotes below)[17]: This publicly available dataset contains 43,472 valid free-text echo reports from the intensive care unit at the Beth Israel Deaconess Medical Center between 2001 and 2012. Similarly, the Findings section of a random subset (n=60) was annotated and used for external validation. Common abbreviations in the text were expanded to their full forms.

Dense Named Entity and Relation Annotation

The annotation schema covers both named entities and relations, as detailed in **Table 1** with corresponding examples. In brief, named entities include "Observation- Definitely Present" (OBS-DP), "Observation- Definitely Absent" (OBS-DA), "Observation- Uncertain" (OBS-UC), "Anatomy" (ANAT), and "Measurements" (MEAS), while relations encompass "Modify," "Gauge," "Located at," and "Suggestive of." Relations were only considered within sentences.

While largely adopting the extraction schema of RadGraph[3], we specifically designed "MEAS" and "Gauge" to reflect the requirements for assessment and grading in echo reports (**Figure 1**).



To accommodate the extended length of echo reports while conforming to our schema requirements, our annotation strategy targeted phrase-level entities such as "left ventricle," "left atrium," and "regional wall motion abnormality" instead of individual words, simplifying the annotation process. Statements in the report that were unrelated to diagnosis, such as quality control notes (e.g., "study performed at the bedside"), were classified as irrelevant and not used for training. The labels of irrelevant were then removed to use the corresponding text as negative examples for model training. An open-source annotation platform, Doccano, was used for annotation (https://github.com/doccano/doccano). The raw annotations were used to train the model without further processing.

Table 1. Examples of Entities and Relations	
Category	**Example**
Named Entities	
Observation- Definitely Present (OBS-DP)	Thrombus, echo density.
Observation- Definitely Absent (OBS-DA)	No regional wall motion abnormality.
Observation- Uncertain (OBS-UC)	Possible foramen ovale.
Anatomy (ANAT)	Left ventricle, Right ventricle.
Measurements (MEAS)	Ejection fraction, size, and severity of

	valvular heart diseases.
Irrelevant (IRRE)	The study was performed at the bedside.
Relations	
Modify	"Left ventricle" modifies "size"
Gauge	"55%" gauge "ejection fraction"
Located at	"Mass" located at "left ventricle"
Suggestive of	"Grade 1/3" suggestive of "normal filling pressure"

Annotating Consistency and Interobserver Agreement

As part of the annotation quality control process, the three annotators (FG, AKM, MTA) all independently annotated a small sample (n=10) from the Mayo dataset to evaluate the interobserver agreement, and the consistency of the annotations was reviewed by senior cardiologists (JFM and CJC) and revised according to the group consensus. The agreement, defined as an exact match of annotations for entities (e.g., 3-element list of ["start word index", "end word index", "label"]) and relations (e.g., 5-element list of ["from start word index", "from end word index", "to start word index", "to end word index", "type"]) were averaged over all annotated samples.

Model Training

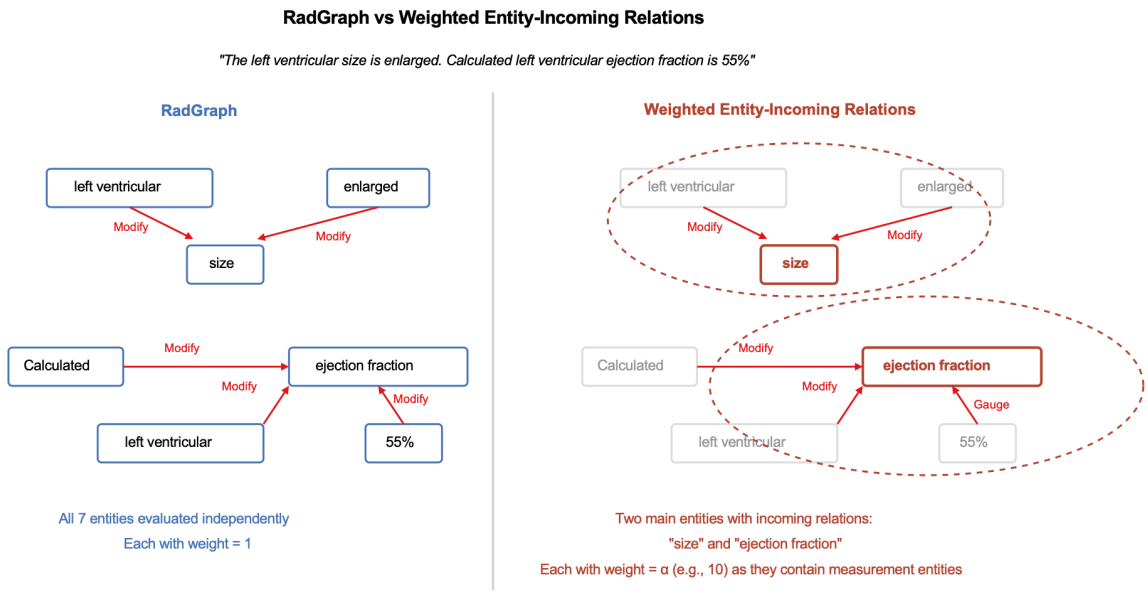
We fine-tuned a pre-trained Bidirectional Encoder Representations from Transformers (BERT)-based model (sciBERT)[18] using the DYGIE++ framework for joint entity and relation extraction[3,6]. The framework was selected due to its effectiveness in similar studies[3,6]. Specifically, the model was trained for 50 epochs using a learning rate of 5e-5, weight decay of 0.01, and batch size of 1 on a 24GB NVIDIA RTX A5000 GPU.

Model Performance Evaluation

We adopted the evaluation approach in the RadGraph study[3] to report the micro F1 for named entity recognition and relation extraction, concerning the imbalanced distribution between classes. For entity prediction, a correct prediction is defined as an exact match of the span boundaries and entity type between the model's prediction and the annotation (e.g., [1, 2, ANAT] matches [1, 2, ANAT]). Similarly, for relation extraction, a correct prediction requires an exact match of the span boundaries of the entity pair and the relation type (e.g., [1, 2, 3, 4, Modify] matches [1, 2, 3, 4, Modify]).

EchoGraph F1 score

An F1 score can be derived from the predictions of the EchoGraph model to automatically assess echo report quality against reference reports. We introduced a new F1-style reward, "weighted entity-incoming relations reward", for EchoGraph's reward calculation. This reward adopts a node-centric subgraph approach by defining "central entities", which are entities without outgoing relations, and gives more weight to other entities with direct relations to the main entities (**Figure 2**).



To emphasize the critical role of numeric measurements, these enriched semantic units are assigned higher weights if the central entity itself or any entity with incoming relations has a numeric measurement (label: MEAS).
Formally, each enriched entity e is assigned a weight $w(e)$:

$$w(e) = \begin{cases} \alpha, & \text{if } e \text{ or any inward-related entity has label MEAS} \\ 1, & \text{otherwise} \end{cases}$$

where $\alpha > 1$ (10 was used in our experiments), thereby prioritizing the accuracy of numeric measurements. Precision and recall are computed using these weighted counts:

$$\text{Precision} = \frac{\sum_{e \in \mathcal{H} \cap \mathcal{R}} w(e)}{\sum_{e \in \mathcal{H}} w(e)}, \quad \text{Recall} = \frac{\sum_{e \in \mathcal{H} \cap \mathcal{R}} w(e)}{\sum_{e \in \mathcal{R}} w(e)}$$

The weighted F1-score is then:

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

<u>The Impact of Data Corruption on Evaluation Metrics in Echo Reports</u>

We examined how automated evaluation metrics respond to corrupted data in echo reports. Using a dataset of 1,000 Mayo Clinic echo reports from Arizona[16], we systematically replaced measurement values and directional terms ("left"/"right") at varying degrees (25%, 50%, and 100%). We then analyzed

how these manipulations affected Echograph F1 and RadGraph F1 scores when comparing the altered reports against their original versions.

Error Analysis

An error analysis was conducted to identify whether the source of the error originated from the entity span or the prediction label, or was completely missing. The definition of mismatch categories is provided in **Supplemental Table 1.**

Results

Dataset: Demographics and baseline diseases

The Mayo dataset comprises 600 patients with a mean age of 64.9 ± 15.8 years, predominantly white (91.8%), and 54.1% male. A majority of the studies (83.8%) were TTE. Baseline comorbidities are detailed in **Supplemental Table 2**. The MIMIC-EchoNotes dataset also had 83.3% TTE studies, but demographic and comorbidity information was unavailable for the anonymized data[19].

Dataset: Annotation

The Mayo dataset has a total of 48,256 entities and 29,731 relations across all 600 reports, and the MIMIC-EchoNotes has a total of 3,672 entities and 2,360 relations across the 60 reports. The most predominant entity category was 'Anatomy' (30.3%) and 'Observation- definitely present' (39.5%) for the Mayo and EchoNotes datasets, respectively. The frequency of 'Measurement' is lower in EchoNotes compared to the reports at Mayo (16.8% vs. 29.3%). Relatively low frequencies of 'Observation-definitely absent', 'Observation-uncertain,' 'Located at,' and 'Suggestive of' were observed. Notably, there were no 'OBS-UC' examples in our Mayo test set, as determined by our patient-level random split procedure. Details are summarized in **Table 2**.

Table 2: Annotation Counts

	Mayo (n=600)	**MIMIC-EchoNotes (n=60)**		**Mayo**	**MIMIC-EchoNotes**
Entity	**Count (%)**		**Relation**	**Count(%)**	
Total	48256 (100%)	3672 (100%)	**Total**	29731 (100%)	2360 (100%)
Anatomy	14606 (30.3%)	1287 (35.0%)	**Gauge**	3345 (11.3%)	84 (3.6%)
Measurement	14146 (29.3%)	618 (16.8%)	**Modify**	25625 (86.2%)	2156 (91.4%)
Observation- definitely present	14122 (29.3%)	1449 (39.5%)	**Located at**	597 (2.0%)	92 (3.9%)
Observation- definitely absent	3118 (6.5%)	249 (6.8%)	**Suggestive of**	164 (0.6%)	28 (1.2%)
Observation- uncertain	28 (0.1%)	10 (0.3%)			
Irrelevant	2236 (4.6%)	59 (1.6%)			

Schema Coverage

Our information extraction schema resulted in a high percentage of schema coverage. In the Mayo data set, the percentage of annotated tokens (among the whole report tokens) was 79.8 ± 3.9%, and 73.6 ± 7.1% in the MIMIC-EchoNotes dataset, as summarized in **Table 3**.

Table 3. Schema Coverage

	Mayo	**MIMIC-EchoNotes**
Dataset	n=600	n=60
Tokens per Report	216.3 ± 54.2	155.2 ± 43.0
Sentence Length	8.1 ± 5.6	7.8 ± 6.8
Annotated Tokens per Report	172.3 ± 42.2	113.5 ± 30.9
Percentage of Annotated Tokens	79.8 ± 3.9%	73.6 ± 7.1%

*Text was transformed into tokens with the NLTK tokenizer

Interobserver Agreement

The initial annotation yielded an average agreement of 0.82 for entities and 0.60 for relations. After review and consensus among the three annotators, the average agreement improved to 0.91 for entities and 0.71 for relations.

Model Performance

The EchoGraph model demonstrated robust performance on the Mayo test set for both entity and relation extraction tasks, while also showing generalizability when evaluated on the external MIMIC-EchoNotes dataset (**Table 4**). When examined by individual entities and relations, EchoGraph demonstrated the best performance in identifying Measurement and Observation: Definitely Present on the Mayo dataset. However, its performance on Measurement and Observation: Definitely Absent decreased when evaluated on the MIMIC-EchoNotes dataset. In terms of relations, our model performed well on more common categories (Gauge, Modify, and Located at) while showing lower performance on the "Suggestive of" category. A more significant performance drop of "Located at" and failing in predicting "Suggestive of" were observed when evaluating on the MIMIC-EchoNotes dataset (**Table 5**).

Table 4. EchoGraph's Performance

		Train	Val	Test	MIMIC-EchoNotes
Precision (PPV)	Entity	0.98	0.83	0.86	0.82
Recall (Sensitivity)	Entity	0.98	0.81	0.84	0.78
Micro F1 Score	Entity	0.98	0.82	0.85	0.80
Precision (PPV)	Relation	0.98	0.85	0.72	0.64
Recall (Sensitivity)	Relation	0.97	0.84	0.69	0.50
Micro F1 Score	Relation	0.98	0.84	0.70	0.56

PPV: positive predictive value

Tabel 5. Performance by entities and relations

	Entities					Relations			
Mayo									
	ANAT	**MEAS**	**OBS-DP**	**OBS-DA**	***OBS-UC**	**Located at**	**Modify**	**Gauge**	**Suggestive of**
Precision	0.83	0.86	0.89	0.82	--	0.73	0.71	0.79	0.40
Recall	0.82	0.86	0.85	0.83	--	0.70	0.68	0.77	0.33
Micro F1	0.82	0.86	0.87	0.82	--	0.71	0.70	0.78	0.36
MIMIC-EchoNotes									
Precision	0.85	0.69	0.87	0.67	0.0	0.42	0.64	0.60	0.0
Recall	0.80	0.69	0.82	0.66	0.0	0.17	0.51	0.43	0.0
Micro F1	0.82	0.69	0.85	0.67	0.0	0.24	0.57	0.50	0.0

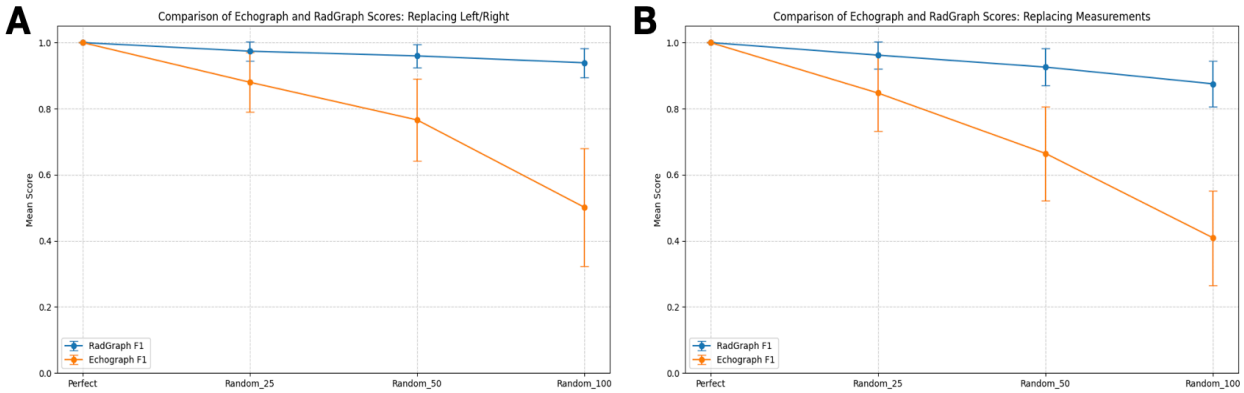
ANAT: Anatomy; MEAS: Measurement; OBS-DP: Observation-Definitely Present; OBS-DA: Observation-Definitely Absent; OBS-UC: Observation-Uncertain. *There were no OBS-UC examples in the Mayo test split.

Error Analysis

Error analysis revealed that span mismatch was the predominant error type across all categories of entities (54.9%) and relations (98.2%) on the Mayo dataset (**Supplemental Table 3**). On the MIMIC-EchoNotes dataset, missing entities constituted the primary error type (47.7%), followed by span mismatch (36.1%), while span mismatch remained the dominant error in relation detection (92.8%) (**Supplemental Table 4**).

EchoGraph F1's Performance on Corrupted Synthetic Echo Reports

Both EchoGraph F1 and RadGraph F1 achieved perfect scores (1.0) when comparing unaltered reports. EchoGraph F1 demonstrated sensitivity to synthetic reports with 25%, 50%, and 100% corrupted modifiers or numbers (50% and 60% drop on completely corrupted reports, respectively). In contrast, RadGraph F1 showed minimal sensitivity to the corruption in directional terms and measurement numbers, with only modest score reductions (6% and 12% drop, respectively) on completely corrupted reports (**Figure 3**).



Discussion

The contributions of this work are: (1) developing EchoGraph, a BERT-based model with a specialized schema for extracting entities and relations (particularly measurements) from echo reports, and (2) implementing a novel F1-style evaluation metric that prioritizes clinically significant entities and relations for echo reports. EchoGraph overcomes limitations of previous approaches by providing a domain-specific automatic assessment tool for echo report quality.

Pros and Cons of Expanded Schema

Achieving micro F1 scores of 0.85 and 0.70 in entity and relation prediction tasks, EchoGraph's performance is close but slightly lower than RadGraph and RadGraph-XL's (**Tables 4 and 5**)[3,6]. We attribute these results primarily to our expanded extraction schema and the use of phrase-level rather than word-level entity spans. The addition of the measurement category potentially confused the model, as these elements were previously classified under the observation category. The use of phrase-level entity spans was also identified as a common source of disagreement in our annotation quality assessment. For instance, the phrase "calculated 2D-linear" (which modifies ejection fraction) could be annotated as either a single entity or as two separate measurement entities. It is also anticipated that annotations on longer reports are more subject to inter-observer variability[3,6]. Similarly, in our error analysis, we found that a majority of the errors stemmed from the entity span (**Supplemental Table 3**), despite correct entity type predictions.

Overall, EchoGraph demonstrated generalizable performance on the MIMIC dataset, despite dropped performance in low-frequency categories (**Tables 4 and 5**). These low-frequency categories, including "OBS-UC" entity, "Located at" and "Suggestive of" relations, were also found to be low-frequency on

reports from other modalities[6]. Notably, the shorter report and sentence lengths in the MIMIC-EchoNotes dataset pose challenges for relation annotation and detection, as denser annotations are required in limited space.

An automatic factual correctness metric for AI-generated echo reports

Established natural language processing metrics such as BLEU, ROUGE, and METEOR[20–22] were not originally designed for evaluating medical texts[16,23]. While RadGraph has provided a strong foundation for the evaluation of radiology reports, its scoring mechanism has limitations when applied to domains like echo reports[16], where numerical precision and relational semantics are critical.

RadGraph's uniform treatment of all entity types—including anatomical terms, modifiers, and numeric values—fails to account for the disproportionate clinical significance of certain measurements in echo. Furthermore, RadGraph's independent treatment of relations disregards the semantic dependency structure between entities, enabling spurious matches when core numeric content is inaccurate (**Figure 3**)[16]. In contrast, by incorporating a category specifically tailored for measurements and corresponding relations in echo reports, EchoGraph more accurately reflects the semantic coherence and numeric precision essential for clinically meaningful echo reporting.

Similar to its predecessor models[3,6], EchoGraph can be leveraged for similar tasks, including automatic annotation of large-scale unstructured echo report text, to evaluate and improve the quality of AI-generated echo reports[3,6,16], as well as enhancing vision-language models in identifying fine-grained pairs between images and corresponding text[24].

EchoGraph is subject to limitations such as dependence on text order and exact text match. For example, "no significant valvular diseases" is equivalent to a ground truth of "trivial mitral and tricuspid regurgitation," yet this could be rated low. Such issues could be addressed by language models that understand the clinical meaning of these equivalent statements with specific fine-tuning[7]. As a BERT-based model, EchoGraph is smaller and faster than large language models[14,15,25] , requiring fewer computational resources as a cost-effective solution for report quality assessments. The development and external validation datasets were both from tertiary referral medical centers and, therefore, may not be representative of a more general population. Although phrase-level entity annotation introduced variation that impacted the model's performance on external datasets, this approach was chosen for its feasibility, considering the relatively long context of the reports.

Conclusion

EchoGraph advances automatic echo report evaluation through its specialized entity-relation extraction schema and clinically weighted rewards that prioritize measurements. This approach provides an effective and scalable solution for evaluating and enhancing the development of language model-based applications in echo.

Data Availability

We released our annotations on the external MIMIC-EchoNote dataset, as well as EchoGraph-annotated MIMIC-EchoNote data to facilitate future research. The Mayo Clinic data set is not publicly available due

to patient privacy considerations, however, it can be obtained from the corresponding author upon reasonable request.

Code Availability

The EchoGraph model is released at: https://github.com/Stanford-AIMI/radgraph

Author Contributions:

C.J.C.: Conceptualization, methodology, software, validation, data curation, formal analysis, writing – original draft, visualization.
J.B.D.: Methodology, software.
M.A.: Methodology, software.
I.B.: Methodology, resources, writing – review and editing.
J.M.F.: Data curation.
F.G.: Data curation.
A.K.M.: Data curation.
M.T.A.: Data curation.
R.A.: Resources, writing – review and editing.
G.C.K.: Resources, writing – review and editing.
J.K.O.: Resources, writing – review and editing.
B.J.E.: Resources, writing – review and editing.
F.F.L.: Methodology, resources, writing – review and editing, supervision.
E.A.: Conceptualization, methodology, resources, writing – review and editing.
C.P.L.: Conceptualization, resources, writing – review and editing, supervision.
All authors have read and approved the manuscript.

Figure Legend

Figure 1. A depiction of the text annotation of an echo report and the corresponding knowledge graph. **Panel A** illustrates the annotation of two sentences: "left ventricular ejection fraction 55%" and "no regional wall motion abnormality." Here, "left" and "ventricular" are annotated as anatomy entities, while "ejection fraction" and "55%" are annotated as measurement entities. "No regional wall motion abnormality" is annotated as Observation: Definitely Absent. The relations between entities are marked as "Modify" and "Gauge." **Panel B** shows the corresponding knowledge graph for this annotation. Notably, "regional wall motion abnormality" is isolated from other entities as no relations are annotated between the two sentences.

Figure 2. Illustration comparing RadGraph's independent entity evaluation versus EchoGraph's weighted entity-relation reward approach. Using a sample sentence describing left ventricular size and ejection fraction measurements, RadGraph evaluates detected entities independently without additional weighting. In contrast, EchoGraph identifies main entities (typically measurements) and assigns a desired weight alpha (here alpha = 10) to their incoming relations during reward calculation.

Figure 3. Both RadGraph F1 and EchoGraph F1 achieved perfect scores (1.0) on unaltered reports. **Panel A**: Performance on reports with randomly replaced left and right modifiers. EchoGraph F1 showed high sensitivity to corrupted modifiers (50% drop) while RadGraph F1 maintained near-perfect scores with minimal sensitivity (6% drop). **Panel B**: Performance on reports with randomly replaced measurements. EchoGraph F1 demonstrated significant sensitivity to number corruption (60% drop) while RadGraph F1 showed only modest reduction (12% drop).

Supplemental Materials

Supplemental Table 1. Definition of Mismatch Categories		
Mismatch Categories	**Definition (Entity)**	**Definition (Relations)**
Span	Incorrect starting OR ending word index	Incorrect starting OR ending entity span
Type	Incorrect label	Incorrect label
Span and Type	Incorrect starting OR ending word index, AND incorrect label	Incorrect starting OR ending entity span, AND incorrect label
Missing	Incorrect starting AND ending word index AND incorrect label	Incorrect starting AND ending entity span AND incorrect label

Supplemental Table 2. Patient Demographics		
	Mayo	**MIMIC-EchoNote**
	n=600	n=60
Age	64.9 ± 15.8	–
Race		
White	551 (91.8%)	–
Black	19 (3.2%)	–
Other	20 (3.3%)	–
Asian	6 (1.0%)	–
Native American	3 (0.5%)	–
Pacific Islander	1 (0.2%)	–
Sex		
Male	324 (54.1%)	–
Female	275 (45.9%)	–
HTN	82 (13.7%)	–
DM	100 (16.7%)	–
CAD	49 (8.2%)	–
CKD	104 (17.3%)	–

Stroke	24 (4.0%)	–
Echo Study Type		
Adult TTE	502 (83.8%)	50 (83.3%)
Adult TEE	46 (7.7%)	10 (16.7%)
Exercise Stress	31 (5.2%)	–
Pharmacological Stress	20 (3.3%)	–

Supplemental Table 3. Error Analysis on the Mayo Dataset

Issue Type	ANAT	MEAS	OBS-DA	OBS-DP	OBS-UC	Total
Entity						
Missing	38 (13.5%)	38 (18.4%)	5 (8.8%)	49 (21.9%)	–	130(16.9%)
Span and type mismatch	115 (40.8%)	13 (6.3%)	17 (29.8%)	47 (21.0%)	–	192(24.9%)
Span mismatch	128 (45.4%)	150 (72.5%)	34 (59.6%)	111 (49.6%)	–	423(54.9%)
Type mismatch	1 (0.4%)	6 (2.9%)	1 (1.8%)	17 (7.6%)	–	25(3.2%)
Relation						
	Gauge	**Located at**	**Modify**	**Suggestive of**		**Total**
Missing	6 (7.6%)	4 (26.7%)	0 (0.0%)	3 (25.0%)	–	13 (1.4%)
Span and type mismatch	0 (0.0%)	1 (6.7%)	0 (0.0%)	1 (8.3%)	–	2 (0.2%)
Span mismatch	73 (92.4%)	10 (66.7%)	851 (99.8%)	8 (66.7%)	–	942 (98.2%)
Type mismatch	0 (0.0%)	0 (0.0%)	2 (0.2%)	0 (0.0%)	–	2 (0.2%)

ANAT: Anatomy; MEAS: Measurement; OBS-DP: Observation-Definitely Present; OBS-DA: Observation-Definitely Absent; OBS-UC: Observation-Uncertain.

Supplemental Table 4. Error Analysis on EchoNote-MIMIC Dataset

Issue Type	ANAT	MEAS	OBS-DA	OBS-DP	OBS-UC	Total
Entity						
Missing	157 (57.7%)	91 (44.0%)	17 (18.5%)	138 (50.2%)	5 (50.0%)	408 (47.7%)
Span and type mismatch	37 (13.6%)	6 (2.9%)	38 (41.3%)	11 (4.0%)	0 (0.0%)	92 (10.7%)
Span mismatch	78 (28.7%)	94 (45.4%)	32 (34.8%)	105 (38.2%)	0 (0.0%)	309 (36.1%)
Type mismatch	0 (0.0%)	16 (7.7%)	5 (5.4%)	21 (7.6%)	5 (50.0%)	47 (5.5%)
Relation						
	Gauge	**Located at**	**Modify**	**Suggestive of**		**Total**
Missing	6 (12.8%)	36 (47.4%)	0 (0.0%)	14 (50.0%)	–	56 (4.7%)
Span and type mismatch	3 (6.4%)	5 (6.6%)	0 (0.0%)	14 (50.0%)	–	22 (1.9%)
Span mismatch	37 (78.7%)	33 (43.4%)	1033 (99.5%)	0 (0.0%)	–	1103 (92.8%)
Type mismatch	1 (2.1%)	2 (2.6%)	5 (0.5%)	0 (0.0%)	–	8 (0.7%)

ANAT: Anatomy; MEAS: Measurement; OBS-DP: Observation-Definitely Present; OBS-DA: Observation-Definitely Absent; OBS-UC: Observation-Uncertain.

Reference:

1. Gardin JM, Adams DB, Douglas PS, Feigenbaum H, Forst DH, Fraser AG, Grayburn PA, Katz AS, Keller AM, Kerber RE, Khandheria BK, Klein AL, Lang RM, Pierard LA, Quinones MA, Schnittger I, Echocardiography AS of. Recommendations for a standardized report for adult transthoracic echocardiography: A report from the American Society of Echocardiography's Nomenclature and Standards Committee and Task Force for a Standardized Echocardiography Report. *J Am Soc*

Echocardiogr 2002;**15**:275–290.
2. Douglas PS, Carabello BA, Lang RM, Lopez L, Pellikka PA, Picard MH, Thomas JD, Varghese P, Wang TY, Weissman NJ, Wilgus R. 2019 ACC/AHA/ASE Key Data Elements and Definitions for Transthoracic Echocardiography: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Data Standards (Writing Committee to Develop Cardiovascular Endpoints Data Standards) and the American Society of Echocardiography. *Circ: Cardiovasc Imaging* 2019;**12**:e000027.
3. Jain S, Agrawal A, Saporta A, Truong SQ, Duong DN, Bui T, Chambon P, Zhang Y, Lungren MP, Ng AY, Langlotz CP, Rajpurkar P. RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. *arXiv* 2021.
4. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt J-N, Laleh NG, Löffler CML, Schwarzkopf S-C, Unger M, Veldhuizen GP, Wagner SJ, Kather JN. The future landscape of large language models in medicine. *Commun Med* 202**3**;3:141.
5. Busch F, Hoffmann L, Rueger C, Dijk EH van, Kader R, Ortiz-Prado E, Makowski MR, Saba L, Hadamitzky M, Kather JN, Truhn D, Cuocolo R, Adams LC, Bressem KK. Current applications and challenges in large language models for patient care: a systematic review. Commun Med 2025;5:26.
6. Delbrouck J-B, Chambon P, Chen Z, Varma M, Johnston A, Blankemeier L, Veen DV, Bui T, Truong S, Langlotz C. RadGraph-XL: A Large-Scale Expert-Annotated Dataset for Entity and Relation Extraction from Radiology Reports. *Find Assoc Comput Linguistics ACL 2024* 2024:12902–12915.
7. Ostmeier S, Xu J, Chen Z, Varma M, Blankemeier L, Bluethgen C, Michalson AE, Moseley M, Langlotz C, Chaudhari AS, Delbrouck J-B. GREEN: Generative Radiology Report Evaluation and Error Notation. arXiv 2024.
8. Liu L, Yang X, Li F, Chi C, Shen Y, Zhang SLM, Ma X, Lyu X, Ma L, Zhang Z, Xue W, Huang Y, Gu J. Towards Automatic Evaluation for LLMs' Clinical Capabilities: Metric, Data, and Algorithm. arXiv 2024.
9. Wu K, Wu E, Wei K, Zhang A, Casasola A, Nguyen T, Riantawan S, Shi P, Ho D, Zou J. An automated framework for assessing how well LLMs cite relevant medical references. *Nat Commun* 2025;**16**:3615.
10. Abbasian M, Khatibi E, Azimi I, Oniani D, Abad ZSH, Thieme A, Sriram R, Yang Z, Wang Y, Lin B, Gevaert O, Li L-J, Jain R, Rahmani AM. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. *npj Digit Med* 2024;**7**:82.
11. Szekér S, Fogarassy G, Vathy-Fogarassy Á. A general text mining method to extract echocardiography measurement results from echocardiography documents. *Artif Intell Med* 2023;**143**:102584.
12. Toepfer M, Corovic H, Fette G, Klügl P, Störk S, Puppe F. Fine-grained information extraction from German transthoracic echocardiography reports. *BMC Méd Inform Decis Mak* 20**15**;15:91.
13. Sun D, Oliveira LM, Spencer KT. Extracting Key Findings Compared in an Echocardiogram Report. *2018 IEEE Int Conf Healthc Inform (ICHI)* 2018:382–383.
14. Yan B, Liu R, Kuo DE, Adithan S, Reis EP, Kwak S, Venugopal VK, O'Connell CP, Saenz A, Rajpurkar P, Moor M. Style-Aware Radiology Report Generation with RadGraph and Few-Shot Prompting. arXiv 2023.
15. Yu F, Endo M, Krishnan R, Pan I, Tsai A, Reis EP, Fonseca EKUN, Lee HMH, Abad ZSH, Ng AY, Langlotz CP, Venugopal VK, Rajpurkar P. Evaluating progress in automatic chest X-ray radiology report generation. *Patterns* 2023;**4**:100802.

16. Chao C-J, Banerjee I, Arsanjani R, Ayoub C, Tseng A, Delbrouck J-B, Kane GC, Lopez-Jimenez F, Attia Z, Oh JK, Erickson B, Fei-Fei L, Adeli E, Langlotz C. Evaluating large language models in echocardiography reporting: opportunities and challenges. *Eur Hear J - Digit Heal* 2025:ztae086.
17. Kwak G, Moukheiber D, Moukheiber M, Moukheiber L, Moukheiber S, Butala N, Celi L, Chen C. EchoNotes Structured Database derived from MIMIC-III (ECHO-NOTE2NUM). *PhysioNet* 2024.
18. Beltagy I, Lo K, Cohan A. SciBERT: A Pretrained Language Model for Scientific Text. arXiv 2019.
19. Gow B, Pollard T, Greenbaum N, Moody B, Johnson A, Herbst E, Waks JW, Eslami P, Chaudhari A, Carbonati T, Berkowitz S, Mark R, Horng S. MIMIC-IV-ECHO: Echocardiogram Matched Subset. PhysioNet 2023.
20. Papineni K, Roukos S, Ward T, Zhu W-J. BLEU: a method for automatic evaluation of machine translation. *Proc 40th Annu Meet Assoc Comput Linguistics - ACL '02* 2002:311–318.
21. Lin C-Y. ROUGE: A Package for Automatic Evaluation of Summaries. Association for Computational Linguistics. p74–81.
22. Lavie A, Agarwal A. Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments. 2007:228–231.
23. Veen DV, Uden CV, Blankemeier L, Delbrouck J-B, Aali A, Bluethgen C, Pareek A, Polacin M, Reis EP, Seehofnerová A, Rohatgi N, Hosamani P, Collins W, Ahuja N, Langlotz CP, Hom J, Gatidis S, Pauly J, Chaudhari AS. Adapted large language models can outperform medical experts in clinical text summarization. *Nat Med* 2024:1–9.
24. Varma M, Delbrouck J-B, Hooper S, Chaudhari A, Langlotz C. ViLLA: Fine-Grained Vision-Language Representation Learning from Real-World Data. arXiv 2023.
25. Zhang X, Wu C, Zhang Y, Xie W, Wang Y. Knowledge-enhanced visual-language pre-training on chest radiology images. Nat Commun 2023;14:4542.