EchoGen: A New Benchmark Study on Generating Conclusions from Echocardiogram Notes

Liyan Tang[1], **Shravan Kooragayalu**[2], **Yanshan Wang**[2],
Ying Ding[1], **Greg Durrett**[1], **Justin F. Rousseau**[1], **Yifan Peng**[3]
[1]The University of Texas at Austin [2]University of Pittsburgh
[3]Weill Cornell Medicine
lytang@utexas.edu, {SHK14, yanshan.wang}@pitt.edu
ying.ding@ischool.utexas.edu, gdurrett@cs.utexas.edu
justin.rousseau@austin.utexas.edu, yip4002@med.cornell.edu

Abstract

Generating a summary from findings has been recently explored (Zhang et al., 2018, 2020) in note types such as radiology reports that typically have short length. In this work, we focus on echocardiogram notes that is longer and more complex compared to previous note types. We formally define the task of echocardiography conclusion generation (**EchoGen**) as generating a conclusion given the findings section, with emphasis on key cardiac findings. To promote the development of EchoGen methods, we present a new benchmark, which consists of two datasets collected from two hospitals. We further compare both standard and state-of-the-art methods on this new benchmark, with an emphasis on factual consistency. To accomplish this, we develop a tool to automatically extract concept-attribute tuples from the text. We then propose an evaluation metric, *FactComp*, to compare concept-attribute tuples between the human reference and generated conclusions. Both automatic and human evaluations show that there is still a significant gap between human-written and machine-generated conclusions on echo reports in terms of factuality and overall quality[1].

1 Introduction

Echocardiography (or echo) is a test that uses sound waves to produce live images of the heart (Mitchell et al., 2019). It has become routinely used to support the diagnosis, management, and follow-up of patients with suspected or known heart diseases. The echo report documents and communicates the evaluation of cardiac and vascular structures in the echocardiography study. As shown in Figure 1, a standard echo report usually consists of a demographic section, an echocardiographic evaluation section (also called the finding section), and a conclusion section (Gardin et al., 2002). In a typical workflow, consultants who interpret echocardiography provide the quantitative measurement and descriptive statements to describe pertinent findings, and then conclude.

In this work, we formally study the task of echo conclusion generation (EchoGen), arising in clinical practice to relieve the clinician of tasks that may contribute to clinician burnout (Alsharqi et al., 2018). A practical system shall be able to generate statements that emphasize abnormal findings, and compare differences and similarities of the current study versus the previous one if available and relevant. We define EchoGen as a task of learning from the demographic and echocardiographic findings section and generating the conclusion section.

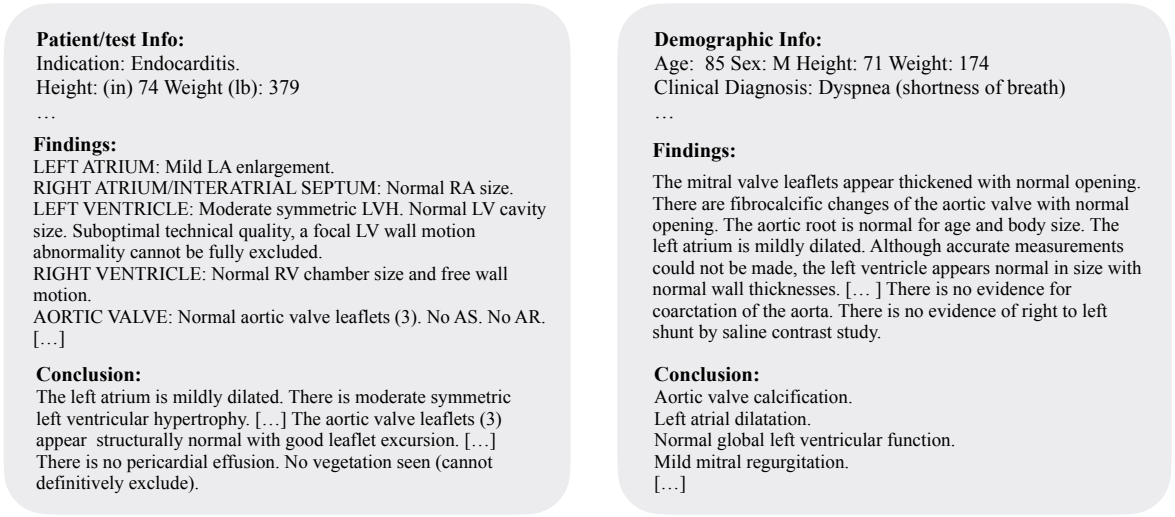
Neural network-based models (See et al., 2017; Lewis et al., 2020) are an attractive method for this task, but are difficult to apply without appropriate training data. To address this gap, we present a large-scale EchoGen benchmark, which consists of two datasets. Here we reply on one prexisting MIMIC-III dataset (EGMIMIC) and one newly collected dataset from the NewYork-Presbyterian Hospital (EGCLEVER) to cover different text genres, data sizes, and degrees of difficulty, and more importantly, highlight common challenges of EchoGen (Figure 1).

Beyond data, a second challenge for EchoGen is to evaluate the factual correctness of a generated conclusion. Automatic metrics such as ROUGE and METEOR only assess content selection but not other quality aspects, such as fluency, grammaticality, and coherence, and are not well-correlated with factuality, leading to the development of separate evaluation measures (Zhang et al., 2018; Falke et al., 2019; Kryscinski et al., 2020; Goyal and Durrett, 2021). This study proposes a new evaluation metric to measure factual consistency, called "FactComp" by considering both concept and their attributes in the fact equivalence criteria.

To better understand the challenge posed by

[1]Code for data construction and model evaluation is available at https://github.com/bionlplab/echo_summarization.

Proceedings of the BioNLP 2022 workshop, Dublin, Ireland, pages 359–368
May 26, 2022.



(a) (b)

Figure 1: Echocardiography reports from the (a) EGMIMIC and (b) EGCLEVER datasets.

EchoGen, we conducted experiments with five baselines: TF-IDF, RANDSENT, LEXRANK, FACTEXT, and BART. We find that BART exceeds other baselines by a large margin, but it has poor transferability when tested on cross-corpus settings. Further human evaluations indicate that there is still a significant gap between generated conclusions and human reference in terms of fluency and factual consistency.

In summary, our contributions can be summarized as follows. (1) We formally introduce the task of EchoGen. (2) We curate a large-scale benchmark from an existing representative dataset and a newly-collected dataset. (3) We introduce a new metric to measure the fact consistency for echo notes. (4) Our metric and human evaluations find that there is still a gap between human reference and generated conclusions for echo reports in terms of fluency and factual consistency.

2 Related works

While EchoGen has not been defined before, there are closely related tasks that were studied before: data-to-text generation, clinical report summarization, and evaluation.

Data-to-text Generation Data-to-text generation is a task of generating text in natural language from non-linguistic input data such as tables and time series (Gatt and Krahmer, 2018; Wiseman et al., 2017; Gardent et al., 2017). Traditional approaches for data-to-text generation (Reiter and Dale, 2000) follow a pipeline of modules such as content selection, text structuring, and surface realization. Recent methods (Gehrmann et al., 2018; Harkous et al., 2020) generate text from data in an end-to-end fashion using the encoder-decoder approach. Data-to-text is also explored in healthcare (Pauws et al., 2019) to facilitate patient review.

Clinical report summarization Clinical report summarization is a long-standing research problem (Adams et al., 2021). Both extractive and abstractive methods have been applied for summarization, covering cases from structured data to text, medical image to text, and history documents to text (Afantenos et al., 2005; Xiong et al., 2019; Pivovarov and Elhadad, 2015).

To the best of our knowledge, few clinical summarization datasets are available. *MEDIQA 2021 ST* provides a task of generating radiology impression statements from textual clinical findings in radiology reports (Ben Abacha et al., 2021) collected from the Indiana University dataset and Stanford Health Care. *CLIP* is a dataset on discharge notes, where the authors' task was to extract the follow-up action items from notes (Mullenbach et al., 2021) . This dataset is more suitable for developing information extraction (IE) systems or extractive summarization methods. Adams et al. (2021) developed a dataset *CLINSUM* from Columbia University Irving Medical Center, focusing on discharge summary notes. While they identified the complex, multi-document summarization task, the dataset is not public to promote the model development by other researchers.

In comparison, our EchoGen is a completely new task on a new note type – echocardiograms. More

importantly, the benchmark covers a diverse range of text genres from two resources. We expect that the models that perform better on both datasets will be more robust in real-word settings.

Evaluation on clinical text Evaluation of clinical text generation or summarization is a challenging research area. Existing methods include automatic approaches and human judgments. For example, commonly used ROUGE-based evaluation metrics measure the overlapping n-grams or longest common sub-sequence between the reference and generated summaries. BERTScore (Zhang et al., 2019) (or HOLMS) is an alternative that accounts for lexical variations by comparing the similarity of semantic representations encoded via BERT (Devlin et al., 2019). However, human evaluations show that these metrics do not always correlate well with factual consistency measurement. Hence, many research works focus on developing automatic consistency metrics that correlate better with human evaluations.

Goodrich et al. (2019) measure the factual consistency as the ratio of overlap between relation triplets under fixed schema extracted from the reference and the generated summary. Kryscinski et al. (2020) propose an entailment-based model FactCC to check whether the source text entails each sentence in the generated summary. Wang et al. (2020) and Durmus et al. (2020) propose QA-based methods that measure the amount of information in the generated summary supported by the source. However, these evaluation approaches often consist of auxiliary modules trained on external or artificial datasets, which is prohibitively expensive and time-consuming to collect. In addition, these modules are hardly generalizable to other clinical settings. Our proposed fact extractor FACTEXT instead relies on linguistic knowledge and is shown to have higher generalizability.

3 EchoGen

3.1 Task definition

We first formulate the EchoGen task. Let $x = \{x_1, ..., x_m\}$ be the demographics and findings section of an echo report, the goal is to generate a conclusion $y = \{y_1, ..., y_n\}$, where m and n are the length of the source section and the generated section of an echo report, respectively. In this work, x is the finding section of a report. We leave leveraging the correlations, if any, between demographic values and generated conclusions into future works.

	EGMIMIC	EGCLEVER
Notes	44,085	13,000
Train	41,164	10,081
Dev	1,447	1,406
Test	1,474	1,513
Source sentences	19	19
Conclusion sentences	14	12
Source tokens	173	219
Conclusion tokens	150	72

Table 1: Statistics for the EchoGen benchmark.

3.2 Dataset construction

The EchoGen benchmark contains two corpora (Table 1. Here, we reply on one prexisting dataset because it is widely used in the clinical NLP community and one newly collected dataset to cover different text styles and levels of difficulties.

EGMIMIC The first dataset was sampled from the MIMIC-III dataset (Medical Information Mart for Intensive Care III) (Johnson et al., 2016). MIMIC-III is a de-identified clinical database composed of over 40,000 patients admitted in the ICUs at Beth Israel Deaconess Medical Center. Of those, we collected echo reports from the `noteevents` table, whose category is "Echo".

We applied the RadText tool [2] to split the notes into a sequence of sections. It uses a rule-based matching algorithm with default rules adapted from SecTag with reported recall of 99% (Denny et al., 2008). We then selected the "Findings" section as the input and the "Conclusion" section as the human reference. We sampled a collection of 41,164, 1,447, and 1,474 reports for training, development, and test, respectively (Table 1). Note that we sampled the echo notes at the patient level. This strategy will ensure that no participant was in more than one group to avoid cross-contamination between the training and test datasets.

EGCLEVER The second dataset is a collection of echo notes in English for heart failure patients from the "Predi*C*tion of Ear*L*y R*E*admissions in Patients with Congesti*VE* Hea*R*t Failure" (CLEVER) cohort at NewYork-Presbyterian Hospital (called EGCLEVER). The patients were admitted and discharged with billing codes ICD-9 Code 428 or ICD-10 Code I50 from January 2008 and July 2018. The study was reviewed and approved by the NewYork-Presbyterian Hospital Institutional Review Board.

[2] https://github.com/bionlplab/radtext

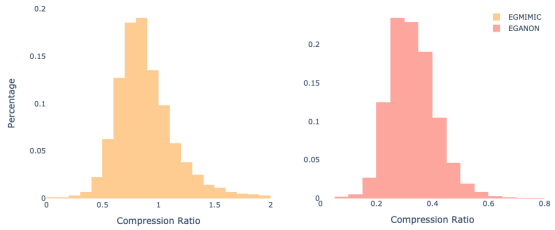


Figure 2: Distribution of word compression ratio on EGMIMIC and EGCLEVER. The ratio defined as the quotient of number of tokens in the reference and that in the source.

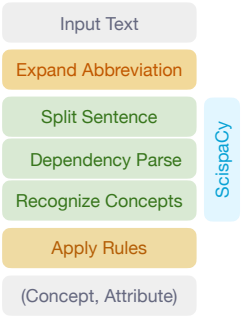


Figure 3: The pipeline of the fact extractor FACTEXT.

We used the same method to preprocess EGCLEVER and sampled a collection of 10,081, 1,406, and 1,513 reports for training, development, and test, respectively.

Comparison The task of EchoGen varies with the data source, which may depend on the individual hospital. Figure 1 shows one echo report from EGMIMIC and one from EGCLEVER. The EGMIMIC report more closely resembles the task of data-to-text generation (Gatt and Krahmer, 2018; Pauws et al., 2019), where the finding section consists of structured data (here, noun phrases in a key-value format), and the conclusion section is written by selecting important findings and expanding them to coherent natural language text. Since data-to-text often has a more complex tabular structure, the result here is somewhere in between pure data and natural language as the tabular structure is not explicit. Therefore, even though the number of tokens in the input is not much shorter than the conclusion section, the conclusion does contain less information than the input.

On the other hand, the conclusion section of our collected dataset EGCLEVER involves more heavily selecting and summarizing content from unstructured text input. The distribution of word compression ratio for both datasets further confirms our observations (Figure 2). The compression ratio is centered around 0.8 for EGMIMIC and 0.3 for EGCLEVER.

3.3 Evaluation Metrics

ROUGE First, we use the standard ROUGE scores (Lin, 2004), and report the F1 scores for ROUGE-1, ROUGE-2, and ROUGE-L, which compare the word-level unigram, bigram, and longest common sequence overlap between the generated and the human reference conclusion, respectively.

Factual Consistency For *Factual Consistency* evaluation, we define a Factual F1 score, inspired by (Zhang et al., 2020). Specifically, we first extract and represent the facts f as a list of "(Concept, Attribute)" pairs $\langle f_1, ...f_n \rangle$. For example, in the sentence "Right ventricular chamber size and free wall motion are normal", the fact list is $\langle$(right ventricular chamber size, normal), (free wall motion, normal)$\rangle$.

The evaluation is then carried out by comparing the list f from the human reference to the list of facts $\hat{f} = \langle \hat{f}_1, ..., \hat{f}_m \rangle$ from a generated conclusion. This requires that a concept and its attributes be extracted correctly to count as one fact.

Finally, the evaluation results are reported using the standard Precision, Recall, and F1-score metrics.

$$P = \frac{1}{|\hat{f}|}\,\mathrm{FE}(f, \hat{f}), \quad R = \frac{1}{|f|}\,\mathrm{FE}(f, \hat{f}),$$
$$F = 2\frac{P \cdot R}{P + R}$$

Here, FE is the factual equivalence criteria and can be defined in various modes.

Strict matching The strict matching mode requires exact matching, and it holds when both the concept and attribute are the same. $\mathrm{FE} = \sum_{\hat{f}_i \in \hat{f}} \sum_{f_j \in f} \mathbb{1}[\hat{f}_i = f_j]$.

BERTScore matching This mode uses greedy matching to maximize the matching similarity. Each fact is matched to the most similar fact in the human reference. Here, we concatenate the attribute with the concept to form a factual noun phrase, and used the BERTScore to measure the similarity between two phrases (Zhang et al., 2019). $\mathrm{FE} = \sum_{\hat{f}_i \in \hat{f}} \max_{f_j \in f} \mathrm{BERTScore}(\hat{f}_i, f_j)$.

	EGMIMIC			EGCLEVER			Overall		
	P	R	F1	P	R	F1	P	R	F1
Findings	94.3	83.4	88.3	88.8	73.1	79.9	91.7	78.5	84.3
Conclusion	91.2	76.1	82.6	96.7	93.5	95.0	93.8	84.4	88.5
Overall	92.8	79.8	85.5	92.8	83.3	87.4	92.8	81.5	86.4

Table 2: The performance of FACTEXT on 25 randomly sampled Echo notes from the validation set of EGMIMIC (13) and EGCLEVER (12). Each report consists of one "Findings" section and one "Conclusion" section. All statistics are obtained by averaging scores from each report.

However, both modes have flaws. For example, strict matching does not consider lexical variation and semantic equivalence. On the other hand, since concept-attribute pairs are supposed to be independent, aligning each fact from the generated conclusion to the most similar one in the reference via BERTScore matching is less meaningful if they are two different facts. Therefore, we relax the definition of these modes and propose approximate matching.

Approximate matching This mode combines strict matching and BERTScore matching. Specifically, a predicted fact is equivalent to a reference fact if their BERTScore is above a threshold t[3]. $\text{FE} = \sum_{\hat{f}_i \in \hat{f}} \sum_{f_j \in f} \mathbb{1}[\text{BERTScore}(\hat{f}_i, f_j) > t]$.

To extract the facts from the text, we develop a rule-based fact extraction system FACTEXT (Figure 3). The tool first splits the text into sentences, and then obtains the universal dependencies (de Marneffe et al., 2021) from the sentences. It further detects UMLS© concepts mentioned in the sentence. Here we focused on the common 55 concepts in the echo notes identified in the data driven way[4]. We used the ScispaCy model (Neumann et al., 2019) trained on MedMentions (Mohan and Li, 2018) to process the text.

Afterward, we applied rules to all identified concepts and subsequently found the attributes that describe the concept. We include negation as an attribute but not uncertainty words as they rarely show up in the text. In this work, we utilized the universal dependency graph to define rules (Chambers et al., 2007). Therefore, the rules take advantage of linguistic knowledge so that the search of attributes is not limited to fixed word distance. The comprehensive rules can be found at our released code. The performance of FACTEXT is discussed in Section 4.

[3]We set threshold $t = 0.85$ in this study based on the performance on the validation set.

[4]Specific concepts are shown in Appendix A.

3.4 Baseline models for benchmarking

We consider 5 baseline models.

TF-IDF Given a source x, TF-IDF first searches for the most similar source x' over all training data based on TF-IDF features and then chooses corresponding conclusion y' as a conclusion for the source x.

RANDSENT We randomly select $k = 12$ sentences from a source as its conclusion, where k is determined according to the average number of conclusion sentences in two collected datasets.

LEXRANK LexRank constructs a graph representation of the course, where nodes are sentences and edges are similarities between sentences (Erkan and Radev, 2004). It then applies the PageRank algorithm on the graph to extract top $k = 12$ most relevant sentences from the source.

FACTEXT We first extract all facts f from a source and then construct a conclusion by concatenating them together. We next convert (Concept, Attribute) pairs into noun phrases by attaching attributes to the beginning of concepts. For example. (right ventricular chamber size, normal) converts to "normal right ventricular chamber size".

BART BART (Lewis et al., 2020) is a pretrained language model that recently demonstrates the state-of-the-art performance in text summarization. It models the conditional likelihood $p(y|x) = \sum_t p(y_t|y_{<t}, x)$, where $y_{<t}$ denotes generated tokens before time step t. We fine-tune a pretrained BART initialized with `facebook/bart-large-xsum` on both datasets.

4 Benchmark results and discussion

Rule-based system Table 2 shows the performance of FACTEXT on randomly sampled 25 ex-

	ROUGE-1			ROUGE-2			ROUGE-L			FC		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
TF-IDF	47.7	47.2	44.9	27.4	27.1	25.9	39.2	38.7	36.9	40.2	41.0	38.8
RANDSENT	58.3	49.2	51.4	34.0	29.7	30.5	47.9	41.0	42.6	49.6	45.8	45.9
LEXRANK	60.5	51.5	53.8	37.0	32.3	33.3	49.9	43.1	44.7	53.6	47.5	48.3
FACTEXT	**69.1**	51.7	57.4	40.0	30.0	33.2	63.8	47.6	52.9	48.8	66.0	54.9
BART	65.5	**67.4**	**69.5**	**55.5**	**57.2**	**55.5**	**65.5**	**67.4**	**65.5**	**72.0**	**66.4**	**67.9**

Table 3: Results on EGMIMIC. ROUGE-1/2/L represent the ROUGE-F1 scores. FC represents Factual Consistency using the approximate matching.

	ROUGE-1			ROUGE-2			ROUGE-L			FC		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
TF-IDF	58.1	57.0	55.5	40.8	40.3	39.1	52.5	51.5	50.2	59.8	60.2	57.8
RANDSENT	37.2	57.7	44.3	17.0	26.9	20.4	28.7	44.9	34.2	33.9	34.3	32.5
LEXRANK	40.2	58.7	46.6	18.0	27.2	21.2	30.8	45.5	35.8	33.4	36.5	33.1
FACTEXT	49.1	49.7	48.3	25.3	25.9	25.0	47.4	47.9	46.6	35.1	50.6	40.4
BART	**76.1**	**72.4**	**73.3**	**63.5**	**60.5**	**61.2**	**73.0**	**69.5**	**70.4**	**85.8**	**73.4**	**78.3**

Table 4: Results on EGCLEVER. ROUGE-1/2/L represent the ROUGE-F1 scores. FC represents Factual Consistency using the approximate matching.

amples from two datasets. Two authors of the work manually annotated all (Concept, Attribute) tuples of sampled examples for evaluation. We obtain Cohen's kappa $\kappa = 0.81$, which indicates a strong agreement. We observe that the system has high precision in all settings but with a drop in recall. This indicates that most (Concept, Attribute) pairs can be correctly identified with a few pairs missed. Further analysis demonstrates that the "Findings" section in EGMIMIC is more well structured than in EGCLEVER. Therefore, FACTEXT on the former setting achieves higher recall and F1.

Baseline Comparisons Table 3 and 4 show the results of baseline approaches on the EGMIMIC and EGCLEVER datasets.

Overall, BART achieves superior performance over other baselines by a large margin, showing the promising result of using abstractive summarization models.

RANDSENT and LEXRANK have similar performances on both datasets. The result is reasonable because LEXRANK relies on inter-sentence similarity to select sentences, but similarities between conclusion sentences are limited in clinical notes.

The TF-IDF baseline has contrary performance on two datasets. Recall that this approach copies the reference directly from the report with the most similar source in the training data. Since the "Conclusion" section is written as structured noun phrases in EGCLEVER and as complete sentences in EGMIMIC, TF-IDF is more likely to achieve a higher ROUGE score in EGCLEVER, which has fewer lexical variations in the "Conclusion" section.

Information Extraction v.s. Text Summarization To tackle the summarization of echocardiography reports as an information extraction (IE) task, we provide our rule-based fact extractor FACTEXT as a performance lower bound. As shown in Table 3 and 4, the rule-based system falls short of performance in both evaluation metrics. Since FACTEXT concatenates all (Concept, Attribute) pairs as noun phrases to form a generated conclusion section, it fails to distill the key information of the source. Further, since the importance of a concept in one report depends on the overall levels of importance of other concepts, external human annotations are required. However, it is hard to reach a consensus on the importance of concepts between domain experts on our dataset (See Human Evaluation below). Therefore, these annotations are deemed to have limited usability, and an IE model trained on them may not be transferable to other clinical datasets.

Alternatively, machine learning based models

Training corpus	EGMIMIC R-1	EGMIMIC R-2	EGMIMIC R-L	EGMIMIC FC	EGCLEVER R-1	EGCLEVER R-2	EGCLEVER R-L	EGCLEVER FC
EGMIMIC	(69.5)	(55.2)	(65.5)	(67.9)	39.9	13.9	24.2	28.1
EGCLEVER	32.6	14.2	23.9	24.9	(73.1)	(60.8)	(70.2)	(78.3)

Table 5: Cross-corpus results of models trained on EchoMIMIC and EGCLEVER using BART. R-1, R-2, R-L represent the ROUGE-F1 scores. FC represents Factual Consistency using the approximate matching. Numbers in parenthesis indicates the performance of the model on the dataset it trained on.

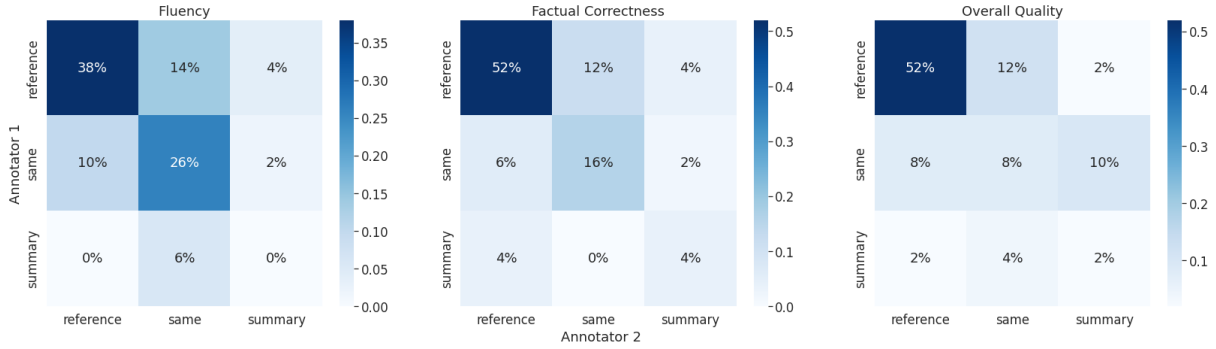


Figure 4: Confusion matrices of human evaluation results on 50 randomly sampled echo notes from EGMIMIC. Results are shown in percentage and "same" means there is a tie between a reference and a generated conclusion.

outperform FACTEXT by a large margin in terms of both ROUGE scores and factual consistency evaluation. This suggests that summarization models can approximate the capability of an IE system and identify more critical facts.

Extractive Summarization v.s. Abstractive Summarization FACTEXT is a strong extractive baseline that selects all concept and attribute pairs f as an extractive conclusion. However, the low recall under our defined evaluation metric indicates that (1) f is not capable of describing all the information in the reference; and (2) domain knowledge is required to generate novel information. The low precision score of FACTEXT, on the other hand, shows that the reference is highly selective of the source text as the majority of facts are excluded from the reference.

Transferability of the model across datasets We intentionally designed the test set to be partially from a hospital system different from the training set (out-of-domain) to test the generalizability of the models. Results are shown in Table 5. As expected, the performance drops significantly in both datasets and is worse than all baselines in Table 3 and 4. The low FC scores indicate that organizations do not share a unified consensus of important information.

5 Human Evaluation

To compare the quality of generated text against a human reference, we conduct a human evaluation following Zhang et al. (2020). We randomly sampled 50 echo reports from the development set of EGMIMIC. For each example, we presented echo findings to two Neurologist and Pulmonary Critical Care physicians along with the human references and summaries generated from BART in random order. We asked the physicians to compare them in three dimensions (1) fluency, (2) factual consistency, and (3) overall quality. For each metric, we asked the physicians to select the better one, with ties allowed.

Since it is difficult to reach an agreement between physicians, we show the human evaluation result as confusion matrices in Figure 4. Across all three dimensions, both physicians agree that human reference is better among half of the selected samples (the upper-left cell of each figure). Further, most of the percentages fall into the top left two-by-two sub-matrices, with the main diagonal being the most frequent. This indicates that physicians have a consensus that generated conclusion is less preferred. There are also uncertainties about whether a reference is better or tied with a generated conclusion (around 20% at off-diagonal). Overall, model-generated summaries are still un-

desired compared to human reference in terms of fluency, factual consistency, and overall quality.

6 Limitations

While our conducted human evaluation suggests that generated summaries from BART tend to have more factual errors than human reference, the accuracy of factuality comparison between BART and other baselines is still limited by the quality of our proposed system FACTEXT. Its performance, especially recall, depends on the accuracy of the ScispaCy model we use and the number of common concepts we focus on (55 in this work). For example, we can integrate the recommended phrases that echocardiographers may choose to use to describe pertinent findings by the American Society of Echocardiography (Gardin et al., 2002). We leave continually designing a more robust information extraction system or learning-based models, which both (1) rely less on domain-specific concepts; and (2) generalize to other types of notes, to future works.

7 Conclusion

In this study, we introduce EchoGen, a new benchmark for evaluating and analyzing models for echocardiography report conclusion generation. We systematically analyze the performance of several baseline methods with our proposed evaluation metric and conclude that there is still a gap between human reference and generated conclusions for echo reports in terms of fluency and factual consistency. Detailed analysis shows that our benchmarking can be used to evaluate the capacity of the models to understand the clinical text and, moreover, to shed light on the future directions for developing clinical text generation and summarization systems.

8 Ethical considerations

The research has been designated by IRB at NewYork-Presbyterian Hospital as Not Human Subject Research. The Protocol Number is 20-10022833.

Acknowledgements

This work was supported by the National Library of Medicine under Award No. 4R00LM013001 and the Amazon Web Services Diagnostic Development Initiative. This work was partially supported by a gift from Amazon and a gift from Salesforce.

References

Griffin Adams, Emily Alsentzer, Mert Ketenci, Jason Zucker, and Noémie Elhadad. 2021. What's in a Summary? Laying the Groundwork for Advances in Hospital-Course Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4794–4811.

Stergos Afantenos, Vangelis Karkaletsis, and Panagiotis Stamatopoulos. 2005. Summarization from medical documents: a survey. *Artificial Intelligence in Medicine*, 33(2):157–177.

M Alsharqi, W J Woodward, J A Mumith, D C Markham, R Upton, and P Leeson. 2018. Artificial intelligence and echocardiography. *Echo Research and Practice*, pages R115–R125.

Asma Ben Abacha, Yassine Mrabet, Yuhao Zhang, Chaitanya Shivade, Curtis Langlotz, and Dina Demner-Fushman. 2021. Overview of the MEDIQA 2021 Shared Task on Summarization in the Medical Domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 74–85, Online. Association for Computational Linguistics.

Nathanael Chambers, Daniel Cer, Trond Grenager, David Hall, Chloe Kiddon, Bill MacCartney, Marie-Catherine De Marneffe, Daniel Ramage, Eric Yeh, and Christopher D. Manning. 2007. Learning alignments and leveraging natural logic. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 165–170.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, pages 1–54.

Joshua Charles Denny, Randolph A. Miller, Kevin B. Johnson, and Anderson Spickard. 2008. Development and evaluation of a clinical note section header terminology. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, pages 156–60.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

Julius M. Gardin, David B. Adams, Pamela S. Douglas, Harvey Feigenbaum, David H. Forst, Alan G. Fraser, Paul A. Grayburn, Alan S. Katz, Andrew M. Keller, Richard E. Kerber, Bijoy K. Khandheria, Allan L. Klein, Roberto M. Lang, Luc A. Pierard, Miguel A. Quinones, Ingela Schnittger, and American Society of Echocardiography. 2002. 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. *Journal of the American Society of Echocardiography: Official Publication of the American Society of Echocardiography*, 15(3):275–290.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *J. Artif. Int. Res.*, 61(1):65–170.

Sebastian Gehrmann, Falcon Dai, Henry Elder, and Alexander Rush. 2018. End-to-end content and plan selection for data-to-text generation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 46–56, Tilburg University, The Netherlands. Association for Computational Linguistics.

Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 166–175, New York, NY, USA. Association for Computing Machinery.

Tanya Goyal and Greg Durrett. 2021. Annotating and Modeling Fine-grained Factuality in Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462, Online. Association for Computational Linguistics.

Hamza Harkous, Isabel Groves, and Amir Saffari. 2020. Have your text and use it too! end-to-end neural data-to-text generation with semantic fidelity. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2410–2424, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-Wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches out: Proceedings of the ACL-04 Workshop*, volume 8, pages 1–8. Barcelona, Spain.

Carol Mitchell, Peter S. Rahko, Lori A. Blauwet, Barry Canaday, Joshua A. Finstuen, Michael C. Foster, Kenneth Horton, Kofo O. Ogunyankin, Richard A. Palma, and Eric J. Velazquez. 2019. Guidelines for Performing a Comprehensive Transthoracic Echocardiographic Examination in Adults: Recommendations from the American Society of Echocardiography. *Journal of the American Society of Echocardiography: Official Publication of the American Society of Echocardiography*, 32(1):1–64.

Sunil Mohan and Donghui Li. 2018. MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts. In *Automated Knowledge Base Construction (AKBC)*.

James Mullenbach, Yada Pruksachatkun, Sean Adler, Jennifer Seale, Jordan Swartz, Greg McKelvey, Hui Dai, Yi Yang, and David Sontag. 2021. CLIP: A Dataset for Extracting Action Items for Physicians from Hospital Discharge Notes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1365–1378, Online. Association for Computational Linguistics.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models

for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Steffen Pauws, Albert Gatt, Emiel Krahmer, and Ehud Reiter. 2019. *Making Effective Use of Healthcare Data Using Data-to-Text Technology*, pages 119–145. Springer International.

Rimma Pivovarov and Noémie Elhadad. 2015. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association: JAMIA*, 22(5):938–947.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Y. Xiong, B. Tang, Q. Chen, X. Wang, and J. Yan. 2019. A study on automatic generation of chinese discharge summary. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1681–1687.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz. 2018. Learning to summarize radiology findings. In *EMNLP 2018 workshop on health text mining and information analysis*.

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. 2020. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

A Echo Concepts

aneurysm, anular, aorta, apex, appendage, arch, arteriosus, artery, atheroma, atrial, atrium, calcification, cava, cavity size, chamber size, chordae, color doppler, defect, disease, effusion, ejection fraction, excursion, foramen, hypertension, hypertrophy, inflammation, leaflet, mitral, muscles, ovale, pad, pericardium, pressure, prolapse, prosthesis, regurgitation, ring, root, septum, shortening, sinus, space, stenosis, structure, tamponade, thicknesses, thrombus, tricuspid, valve, vegetation, velocities, velocity, ventricle, ventricular, wall motion.