AJNR

AI-Assisted Summarization of Radiological Reports: Evaluating GPT3davinci, BARTcnn, LongT5booksum, LEDbooksum, LEDlegal, and LEDclinical

Aichi Chien, Hubert Tang, Bhavita Jagessar, Kai-wei Chang, Nanyun Peng, Kambiz Nael and Noriko Salamon

AJNR Am J Neuroradiol published online 18 January 2024
http://www.ajnr.org/content/early/2024/01/18/ajnr.A8102

Published January 18, 2024 as 10.3174/ajnr.A8102

AI-Assisted Summarization of Radiological Reports: Evaluating GPT3davinci, BARTcnn, LongT5booksum, LEDbooksum, LEDlegal, and LEDclinical

Aichi Chien, Hubert Tang, Bhavita Jagessar, Kai-wei Chang, Nanyun Peng, Kambiz Nael, and Noriko Salamon

ABSTRACT

BACKGROUND AND PURPOSE: The review of clinical reports is an essential part of monitoring disease progression. Synthesizing multiple imaging reports is also important for clinical decisions. It is critical to aggregate information quickly and accurately. Machine learning natural language processing (NLP) models hold promise to address an unmet need for report summarization.

MATERIALS AND METHODS: We evaluated NLP methods to summarize longitudinal aneurysm reports. A total of 137 clinical reports and 100 PubMed case reports were used in this study. Models were 1) compared against expert-generated summary using longitudinal imaging notes collected in our institute and 2) compared using publicly accessible PubMed case reports. Five AI models were used to summarize the clinical reports, and a sixth model, the online GPT3davinci NLP large language model (LLM), was added for the summarization of PubMed case reports. We assessed the summary quality through comparison with expert summaries using quantitative metrics and quality reviews by experts.

RESULTS: In clinical summarization, BARTcnn had the best performance (BERTscore = 0.8371), followed by LongT5Booksum and LEDlegal. In the analysis using PubMed case reports, GPT3davinci demonstrated the best performance, followed by models BARTcnn and then LEDbooksum (BERTscore = 0.894, 0.872, and 0.867, respectively).

CONCLUSIONS: AI NLP summarization models demonstrated great potential in summarizing longitudinal aneurysm reports, though none yet reached the level of quality for clinical usage. We found the online GPT LLM outperformed the others; however, the BARTcnn model is potentially more useful because it can be implemented on-site. Future work to improve summarization, address other types of neuroimaging reports, and develop structured reports may allow NLP models to ease clinical workflow.

ABBREVIATIONS: BART = bidirectional and auto-regressive transformer; BERT = bidirectional encoder representations from transformer; LED = longformer-encoder-decoder; LLM = large language model; NLP = natural language processing; ROUGE = recall-oriented understudy for gisting evaluation

ChatGPT and other large language models (LLMs) have raised interest in using natural language processing (NLP) for clinical research (eg, to collect clinical outcome data).[1-3] We sought to utilize AI technology to improve brain aneurysm research and clinical workflow by summarizing radiology imaging reports.[4] Imaging reports are written by radiologists to describe findings of clinical imaging for disease diagnosis, treatment, and follow-up. As additional imaging is performed and follow-up duration increases, additional time is required to review reports and make the diagnosis in a follow-up visit.[4-6] Longitudinal imaging which captures aneurysm changes over the course of 3–5 years is common.[7,8] This study specifically evaluated AI language models' summarization of a longitudinal series of aneurysm imaging reports. We tested the capability and quality of different NLP models, including a GPT text model. We present methodology that can be used to evaluate NLP models. We aim to promote a systematic approach using quantitative evaluation to understand the performance of NLP models so that clinical researchers can objectively understand the strength and weakness of these new technologies and further harness the benefits they may provide to medical research.

In our analysis using real longitudinal brain aneurysm imaging reports, we first locally implemented 5 state-of-the-art summarization models: BARTcnn (Meta [previously Facebook] Menlo

Received June 28, 2023; accepted after revision November 9.

From the Department of Radiological Science (A.C., H.T., B.J., K.N., N.S.), David Geffen School of Medicine at UCLA, Los Angeles, California; and Department of Computer Science (K.C., N.P.), University of California, Los Angeles, Los Angeles, California.

This work was supported in part by National Institutes of Health R01HL152270.

Please address correspondence to Aichi Chien, PhD, Department of Radiology, David Geffen School of Medicine at UCLA, 10833 LeConte Ave, Box 951721, Los Angeles, CA 90095; e-mail: aichi@ucla.edu; @Chienlab_UCLA; @RadiologyUcla

http://dx.doi.org/10.3174/ajnr.A8102

Park, California),[9] LongT5booksum (Google, Mountain View, California),[10] LEDbooksum, LEDlegal, and LEDclinical (Allen Institute for AI, Seattle, Washington) and compared the performance using expert-generated summaries (ground truth) within the hospital firewall. These are all machine learning models for NLP tasks. These models were developed and shared by Meta, Google, and the Allen Institute for AI, and further improved by individual users for NLP research. For example, BART (bidirectional and auto-regressive transformer [BART]-large-cnn by Meta) was trained on the CNN/DailyMail data set, with >300,000 news articles and their respective summaries from CNN and Daily Mail articles. GPT3davinci (OpenAI, San Francisco), a massive LLM, can only be accessed online and is therefore not a HIPAA-compliant application. But to compare it with the other 5 models, we used 100 publicly accessible English case reports related to brain aneurysms from PubMed.[5] This evaluation allowed us to fairly understand the strengths of all the models in the tasks to generate aneurysm follow-up summarization reports.

MATERIALS AND METHODS

Clinical Collection

This study was approved by the Institutional Review Board. We anonymized and analyzed clinical imaging reports from 52 patients (64 aneurysms) from 2005 to 2022 undergoing monitoring for intracranial aneurysm progression. There were 44 females and 8 males. A total of 137 clinical imaging reports were used for this study. The average interval between the first and second visits, and second and third visits was 15.35 and 12.36 months, respectively. Typically, our center recommends aneurysm follow-up every 12 months. Sometimes patients return later than the recommended length of time. Usually, after the second visit, they will adhere closer to the schedule for the third visit. The aneurysm imaging reports include 3 modalities: MR angiography, CT angiography, and DSA. Online Supplemental Data show the patient demographic and imaging report information.

Case Report Analysis

We also performed comparative analysis and evaluated the NLP summary models using a data set derived from PubMed. A total of 100 publicly accessible English case reports on neurovascular disease (brain aneurysms, neuroangiography, vascular malformations) were collected through PubMed search, and their body texts were used for this part of the analysis. Expert-generated summaries based on these case reports as well as compiled figure captions were used as reference comparisons.

Reference Standards Preparation

H.T., B.J., and A.C., who trained in neuroradiology and brain aneurysm disease clinical reports, performed the following roles: wrote the summaries and graded the PubMed case reports and clinical imaging reports that were used in this research.[11-13] K.N. and N.S., 2 board-certified neuroradiologists, reviewed and revised the summaries and grading. For each patient, summaries were generated from the first visit, first 2 visits, and first 3 visits. Patient information, aneurysm features, imaging technique, and treatment were essential information in the summary.

Summarization Models

We used 6 state-of-the-art NLP summarization models that have been trained with different source data sets and shown strength in faithful summary: BARTcnn—BART model trained on the CNN/DailyMail news data set;[9] LongT5booksum—T5 trained on the booksum data set, a collection of human written summaries of various literature;[10] LEDbooksum—Longformer Encoder-Decoder (LED) model trained on the booksum data set;[14] LEDlegal—LED model trained on sec-litigation-releases data set containing more than 2700 litigation releases and complaints;[14] LEDclinical—LED model trained on references extracted from revised references in the MIMIC-III data set, a database of public health—related data;[15] GPT3davinci—GPT3 davinci-003 LLM utilizing 1.75 billion parameters to generate summaries and only accessible from an API provided by OpenAI.[16] Except for GPT3davinci, models were implemented locally in Python (http://www.python.org). Online Supplemental Data aggregate information about each model and information to access the source code.

Quantitative Evaluation

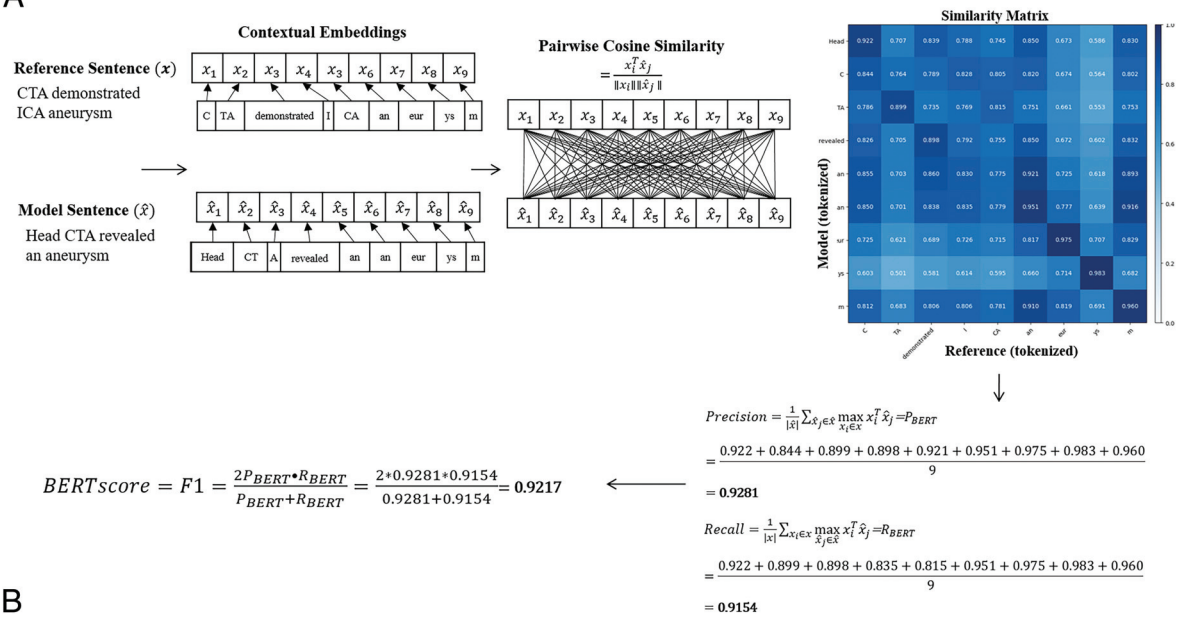
We compared model-outputted summaries (model summary) to expert generated summaries (ground truth summary) through evaluation of recall-oriented understudy for gisting evaluation (ROUGE)-1, ROUGE-2, ROUGE-L score, and BERTscore. ROUGE-1, ROUGE-2, ROUGE-L, and bidirectional encoder representations from transformer (BERT) score are the standard matrix using computational approaches to assess the quality of NLP models.[17,18] We also calculated text reduction: the ratio of the length of the summary with the length of the original text. Online Supplemental Data show the list of equations. Specifically, a unit of text, *grams*, was used in ROUGE-1, ROUGE-2, and ROUGE-L score. (For example, a unigram refers to a single word.) BERTScore evaluates words in the reference summary and model summary by finding the cosine similarity and implements greedy matching techniques to define the score. For each metric, the F1-score, a statistical calculation to estimate the accuracy for each score type, was used to compare the performance of the models. The Figure illustrates the NLP parameters and the processes for calculating these scores.

Expert Evaluation

The expert evaluation was performed by 4 experts (H.T., B.J., K.N., N.S.) based on assessed *readability, accuracy of information, comprehensiveness,* and *redundancy* following the approach proposed by Goldstein et al.[13,19-21] Online Supplemental Data show the evaluation matrix for each category and how the score was evaluated. Specifically, we evaluate each summary by comparing the model-generated summary against the expert generated summary (ground truth) for *readability*—refers to grammatical correctness of each summary, *accuracy*—refers to correctness of information in each summary, *comprehensiveness*—refers to amount of aneurysm information in each summary, and *redundancy*—measures length and amount of redundant information present in each summary. Each summary was evaluated and given a numeric score of 1 to 5 (5 being best) in the categories

	Sentence	ROUGE-1 (Unigram)	ROUGE-2 (Bigram)	ROUGE-L (LCS)
REFERENCE sentence	..at 2 months follow-up, patient recovered completely..	{at, 2, months, follow-up, patient, recovered, completely}	{at 2, 2 months, months follow-up, follow-up patient, patient recovered, recovered completely}	{at, 2, months, follow-up}
MODEL sentence	..the patient made a full recovery at 2 months follow-up..	{the, patient, made, a, full, recovery, at, 2, months, follow-up}	{the patient, patient made, made a, a full, full recovery, recovery at, at 2, 2 months, months follow-up}	
Score		$P = \frac{\#\ unigrams\ in\ REFERENCE\ and\ MODEL}{\#\ unigrams\ in\ MODEL} = \frac{5}{10}$ $R = \frac{\#\ unigrams\ in\ REFERENCE\ and\ MODEL}{\#\ unigrams\ in\ REFERENCE} = \frac{5}{7}$ $F1 = 2\frac{PR}{P+R} = \frac{2*\frac{5}{10}*\frac{5}{7}}{\frac{5}{10}+\frac{5}{7}} = \mathbf{0.588}$	$P = \frac{\#\ bigrams\ in\ REFERENCE\ and\ MODEL}{\#\ unigrams\ in\ MODEL} = \frac{3}{9}$ $R = \frac{\#\ bigrams\ in\ REFERENCE\ and\ MODEL}{\#\ bigrams\ in\ REFERENCE} = \frac{3}{6}$ $F1 = 2\frac{PR}{P+R} = \frac{2*\frac{3}{6}*\frac{3}{9}}{\frac{3}{6}+\frac{3}{9}} = \mathbf{0.400}$	$P = \frac{\#\ unigrams\ in\ LCS}{\#\ unigrams\ in\ MODEL} = \frac{4}{10}$ $R = \frac{\#\ unigrams\ in\ LCS}{\#\ unigrams\ in\ REFERENCE} = \frac{4}{7}$ $F1 = 2\frac{PR}{P+R} = \frac{2*\frac{4}{10}*\frac{4}{7}}{\frac{4}{10}+\frac{4}{7}} = \mathbf{0.471}$

A



$$Precision = \frac{1}{|\hat{x}|}\sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^T \hat{x}_j = P_{BERT}$$

$$= \frac{0.922 + 0.844 + 0.899 + 0.898 + 0.921 + 0.951 + 0.975 + 0.983 + 0.960}{9}$$

$$= \mathbf{0.9281}$$

$$Recall = \frac{1}{|x|}\sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^T \hat{x}_j = R_{BERT}$$

$$= \frac{0.922 + 0.899 + 0.898 + 0.835 + 0.815 + 0.951 + 0.975 + 0.983 + 0.960}{9}$$

$$= \mathbf{0.9154}$$

$$BERTscore = F1 = \frac{2P_{BERT} \bullet R_{BERT}}{P_{BERT} + R_{BERT}} = \frac{2*0.9281*0.9154}{0.9281+0.9154} = \mathbf{0.9217}$$

B

FIGURE. Visualization of quantitative evaluation with examples. The example gives expert generated reference sentence and model-generated sentences. *A*, Different ROUGE scores are calculated based on defining unigrams, bigrams, and longest common subsequences. Matching between the reference and candidate sentences are highlighted in red. ROUGE-1, ROUGE-2, and ROUGE-L are computed as F1 scores using P (precision) and R (recall) values. *B*, BERTscore was calculated by 1) first converting text into tokens, 2) calculating pair-wise cosine similarity between every reference and model token, and 3) identifying the tokens in the other sentence with the highest similarity value, and using the highest similarity values to calculate F1.

independently by H.T. and B.J. The average score was then approved by K.N. and N.S. to be the final score. In the case of discrepancies with a score difference of >3 (medium score), the summary was rescored until consensus was reached and the score was approved.

RESULTS

Our analysis using patient data showed that BARTcnn performed the best overall for single and longitudinal visit reports when both quantitative and expert evaluation was considered (Online Supplemental Data). LEDbooksum ranked second for 1 imaging report, followed by LongT5Booksum. For more than 2 imaging reports, BARTcnn held the best performance followed by LongT5Booksum and LEDlegal. Comparing the reduction of the text, experts were able to reduce report text to an average of 16.29%, 13.14%, and 9.86% for 1, 2, and 3 visit reports, respectively. LEDlegal and LongT5Booksum showed better text reduction but were low in comprehensiveness. BARTcnn provided the next best reduction and maintained reasonable comprehensiveness.

Our comparative study of model performance with GPT3davinci (Online Supplemental Data) found that GPT3davinci demonstrated superior performance in case report summarization, scoring highest in all categories. It was followed by BARTcnn and LEDbooksum. Based on the expert quality analysis, GPT3davinci and BARTcnn had the closest readability, accuracy, and redundancy. BARTcnn, which had the second-best overall performance, could produce comparable output to GPT3davinci while using local workstations (Online Supplemental Data).

DISCUSSION

Given the sophistication of the models and large training sets, we expected GPT3 and BARTcnn to be among the best models for summarization. Although there is room for improvement, these 2 characteristics appear to have been the most important determinants of summarization performance. LEDclinical was trained based on a critical care data base (MIMIC-III), yet the performance was relatively poor.[22] This may be because the radiology reports we summarized are more limited in scope than the broad assortment of clinical notes used to train LEDclinical against

discharge reports, and thus do not benefit from its reference revision approach.[15] Additionally, the radiology reports may not be sufficiently long for the Longformer network model underlying the LED models (LEDclinical, LEDbooksum, LEDlegal) to have an advantage over the other models.[14] The performance of LongT5booksum, LEDbooksum, and LEDlegal on average fell between BARTcnn and LEDclinical. This likely reflects the purpose of the original models (book summarization, and legal document summarization) being more restricted, while GPT3 and BARTcnn were trained on a wider range of topics. However, in some limited contexts (eg, LEDbooksum, single clinical visit quantitative metrics), they were able to surpass BARTcnn. In the expert assessment of redundancy, the BARTcnn model did have a tendency to repeat sentences, but this did not have a large impact on the human readers' ability to parse the summaries. (For example: *Patient will have follow-up MRA in 12 months. Patient will have follow-up MRA in 12 months.*)

Summarization of longitudinal clinical imaging reports is a relatively new area of research. However, in the past, work has focused on both NLP processing of neuroradiology reports[23-25] and extracting temporal relations from clinical reports.[26,27] Processing of neuroradiology reports includes different objectives, such as generating diagnoses or summarization. Recently, neural network NLP models have become the focus, in particular variations of BERT-based models.[28] Models such as ClinicalBERT predict 30-day hospital readmission based on discharge summaries and various intensive care clinical notes, including radiology reports.[29] Recently, a variety of BERT derivatives, collectively referred to as RadBERT, were trained with a broad set of Veterans Affairs' radiology reports and tested on tasks including summarization.[30] The testing methodology differed from that presented here. Summaries generated by experts were not used, and the data set, likewise, was not focused on a specific disease or longitudinal data. Their results therefore are not directly comparable with ours, and their reported ROUGE scores were uniformly lower than our results. The BARTcnn model we tested extends BERT with an autoregressive decoder, effectively adding a key feature of GPT to BERT.[9] The model design, along with initial training set, is a large part of why BARTcnn was able to outperform other local models in our study.

Defining temporal relations in clinical reports is essential to follow the course of disease and treatment. Previous research has largely focused on encoding the sequence of events with a structure digestible by software for further analysis. Based on a few forms of cancer, the clinical data sets (THYME/THYME2) have been carefully annotated.[31] Other diseases and longitudinal series of radiology reports have both not been a focus. Direct summarization via deep learning models, as we present here, has only recently become viable and bypasses the explicit extraction of temporal relationships. This approach shares the pros and cons with many other machine learning solutions. The model is largely a black box that is simple to use and very effective, but because it does not explicitly give structure to the data, modifying the form of the output into, for example, a chart or table may require another model focused on temporal relation extraction.

The models we evaluated showed great potential to help clinical workflow and follow-up management. These models were not designed for clinical imaging reports, and with the exception of LEDclinical, not trained on clinical data sets. Based on our qualitative (expert) evaluation, we identified some common limitations. Online Supplemental Data provide examples and critiques of the output from the different models. Many models had qualitative deductions in readability, including spelling and grammatical errors, as well as a lack of comprehensiveness (Online Supplemental Data). This lack of comprehensiveness often concerned the end of the text, leaving out important information that occurred at the end (eg, patient outcomes or treatment). This limitation means that these existing models are not immediately ready for clinical usage. In the future, better focusing the models on the characteristics of the data set can further improve the models. Future steps to improve summarization can be first to tune the models, adjusting input/output length and other parameters, followed by transfer learning on the longitudinal aneurysm imaging report data set we have now created. Currently, we are researching other related topics as well, including the combination of CNN image-processing with NLP text to process complete imaging reports.[32]

Based on this study, we found that the type of data missing (eg, aneurysm locations) varies between different models. In the current state, the models can only reliably provide information, such as general patient history and sometimes miss critical diagnostic information. However, NLP models continue to rapidly improve. Our intention was to compare these models in a relevant way, so that as new models are developed, we can objectively assess model performance. Comparing different models with expert evaluation can also help computer scientists identify future areas to be addressed and facilitate further developing better AI-generated summaries that can be sufficient for a clinical review.

ACKNOWLEDGMENTS

Due to the sensitive nature of the aneurysm clinical reports collected for this study and to protect patient privacy, requests to access the data by researchers trained in human subject research may be sent to the corresponding author or NIH NHLBI BioLINNC Biologic Specimen and Data Repositories Information Coordinating Center following NIH public access policy reference to project R01HL152270. To access the source code for the models described in this study, links are provided in Online Supplemental Data.

CONCLUSIONS

The AI NLP models showed great potential to generate clinical summaries. Although these models were not created for clinical imaging reports, the summaries were able to capture critical information, albeit not yet at a level suitable for clinical usage. While the GPT model had superior performance, a local BARTcnn model provided comparable quality results. This work showed a new pipeline to evaluate AI NLP models for future neuroimaging report applications. Future steps to improve summarization will be to tune the models with transfer learning on different clinical longitudinal imaging report data sets.

Disclosure forms provided by the authors are available with the full text and PDF of this article at www.ajnr.org.

REFERENCES

1. Shen Y, Heacock L, Elias J, et al. **ChatGPT and other large language models are double-edged swords.** *Radiology* 2023;307:e230163 CrossRef Medline
2. Akinci D'Antonoli T, Stanzione A, Bluethgen C, et al. **Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions**. *Diagn Interv Radiol* 2023 Oct 3. [Epub ahead of print] CrossRef Medline
3. Tippareddy C, Jiang S, Bera K, et al. **Radiology reading room for the future: harnessing the power of large language models like ChatGPT.** *Curr Probl Diagn Radiol* 2023 Aug 30. [Epub ahead of print] CrossRef Medline
4. Moharasan G, Ho TB. **Extraction of temporal information from clinical narratives.** *J Healthc Inform Res* 2019;3:220–44 CrossRef Medline
5. Mishra R, Bian J, Fiszman M, et al. **Text summarization in the biomedical domain: a systematic review of recent research.** *J Biomed Inform* 2014;52:457–67 CrossRef Medline
6. Sarzynski E, Hashmi H, Subramanian J, et al. **Opportunities to improve clinical summaries for patients at hospital discharge.** *BMJ Qual Saf* 2017;26:372–80 Medline
7. Chien A, Callender RA, Yokota H, et al. **Unruptured intracranial aneurysm growth trajectory: occurrence and rate of enlargement in 520 longitudinally followed cases.** *J Neurosurg* 2019;132:1077–87 CrossRef Medline
8. Eskey CJ, Meyers PM, Nguyen TN, American Heart Association Council on Cardiovascular Radiology and Intervention and Stroke Council, et al. **Indications for the performance of intracranial endovascular neurointerventional procedures: a scientific statement from the American Heart Association.** *Circulation* 2018;137:e661–89 CrossRef Medline
9. Lewis M, Liu Y, Goyal N, et al. **BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.** *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020 CrossRef
10. Guo M, Ainslie J, Uthus D, et al. **Longt5: efficient text-to-text transformer for long sequences.** *Findings of the Association for Computational Linguistics: NAACL 2022*. 2022 CrossRef
11. Alfattni G, Peek N, Nenadic G. **Extraction of temporal relations from clinical free text: a systematic review of current approaches.** *J Biomed Inform* 2020;108:103488 CrossRef Medline
12. Bhandari M, Gour PN, Ashfaq A, et al. **Re-evaluating evaluation in text summarization.** *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020 CrossRef
13. Goldstein A, Shahar Y. **An automated knowledge-based textual summarization system for longitudinal, multivariate clinical data.** *J Biomed Inform* 2016;61:159–75 CrossRef Medline
14. Beltagy I, Peters ME, Cohan A. **Longformer: the long-document transformer.** *arXiv* 2020;2004.05150
15. Adams G, Shing HC, Sun Q, et al. **Learning to revise references for faithful summarization.** *arXiv* 2022;2204.10290
16. Goyal T, Jessy Li J, Durrett G. **News summarization and evaluation in the era of gpt-3.** *arXiv* 2022;2209.12356
17. Lin CY. **Rouge: a package for automatic evaluation of summaries.** *Text Summarization Branches Out* 2004:74–81
18. Zhang T, Kishore V, Wu F, et al. **Bertscore: evaluating text generation with Bert.** *arXiv* 2019;1904.09675
19. Hirsch JS, Tanenbaum JS, Lipsky Gorman S, et al. **Harvest, a longitudinal patient record summarizer.** *J Am Med Inform Assoc* 2015;22:263–74 CrossRef Medline
20. Percha B. **Modern clinical text mining: a guide and review.** *Annu Rev Biomed Data Sci* 2021;4:165–87 CrossRef Medline
21. Sun W, Cai Z, Li Y, et al. **Data processing and text mining technologies on electronic medical records: a review.** *J Healthc Eng* 2018;2018:4302425 CrossRef Medline
22. Johnson AE, Pollard TJ, Shen L, et al. **Mimic-III, a freely accessible critical care database.** *Sci Data* 2016;3:160035 CrossRef Medline
23. Fu S, Leung LY, Wang Y, et al. **Natural language processing for the identification of silent brain infarcts from neuroimaging reports.** *JMIR Med Inform* 2019;7:e12109 CrossRef Medline
24. Kim C, Zhu V, Obeid J, et al. **Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroke.** *PLoS One* 2019;14:e0212778 CrossRef Medline
25. Wheater E, Mair G, Sudlow C, et al. **A validated natural language processing algorithm for brain imaging phenotypes from radiology reports in UK electronic health records.** *BMC Med Inform Decis Mak* 2019;19:184 CrossRef Medline
26. Sun W, Rumshisky A, Uzuner O. **Evaluating temporal relations in clinical text: 2012 i2b2 challenge.** *J Am Med Inform Assoc* 2013;20:806–13 CrossRef Medline
27. Moskovitch R, Shahar Y, Wang F, et al. **Temporal biomedical data analytics.** *J Biomed Inform* 2019;90:103092 CrossRef Medline
28. Devlin J, Chang MW, Lee K, et al. **Bert: pre-training of deep bidirectional transformers for language understanding.** *arXiv* 2019;1810.04805
29. Huang K, Altosaar J, Ranganath R. **Clinicalbert: modeling clinical notes and predicting hospital readmission.** *arXiv* 2020;1904.05342
30. Yan A, McAuley J, Lu X, et al. **Radbert: adapting transformer-based language models to radiology.** *Radiol Artif Intell* 2022;4:e210258 CrossRef Medline
31. Styler IW, Bethard S, Finan S, et al. **Temporal annotation in the clinical domain.** *TACL* 2014;2:143–54 CrossRef Medline
32. Monajatipoor M, Rouhsedaghat M, Li LH, et al. **Berthop: an effective vision-and-language model for chest x-ray disease diagnosis.** *Med Image Comput Comput Assist Interv* 2022;13435:725–34 CrossRef Medline