Summarizing Patients' Problems from Hospital Progress Notes Using Pre-trained Sequence-to-Sequence Models

Yanjun Gao*
ICU Data Science Lab
School of Medicine and Public Health
University of Wisconsin-Madison
`ygao@medicine.wisc.edu`

Dmitriy Dligach
Loyola University Chicago
`ddligach@luc.edu`

Timothy Miller, Dongfang Xu
Boston Children's Hospital
and Harvard Medical School
`Timothy.Miller, Dongfang.Xu`
`@childrens.harvard.edu`

Matthew M. Churpek, Majid Afshar
ICU Data Science Lab
School of Medicine and Public Health
University of Wisconsin-Madison
`mchurpek, mafshar@medicine.wisc.edu`

Abstract

Automatically summarizing patients' main problems from daily progress notes using natural language processing methods helps to battle against information and cognitive overload in hospital settings and potentially assists providers with computerized diagnostic decision support. Problem list summarization requires a model to understand, abstract, and generate clinical documentation. In this work, we propose a new NLP task that aims to generate a list of problems in a patient's daily care plan using input from the provider's progress notes during hospitalization. We investigate the performance of T5 and BART, two state-of-the-art seq2seq transformer architectures, in solving this problem. We provide a corpus built on top of progress notes from publicly available electronic health record progress notes in the Medical Information Mart for Intensive Care (MIMIC)-III. T5 and BART are trained on general domain text, and we experiment with a data augmentation method and a domain adaptation pre-training method to increase exposure to medical vocabulary and knowledge. Evaluation methods include ROUGE, BERTScore, cosine similarity on sentence embedding, and F-score on medical concepts. Results show that T5 with domain adaptive pre-training achieves significant performance gains compared to a rule-based system and general domain pre-trained language models, indicating a promising direction for tackling the problem summarization task.

1 Introduction

The progress note is a common note type in the electronic health record (EHR) that also contains the necessary details for medical billing; therefore, every hospital day will contain at least one
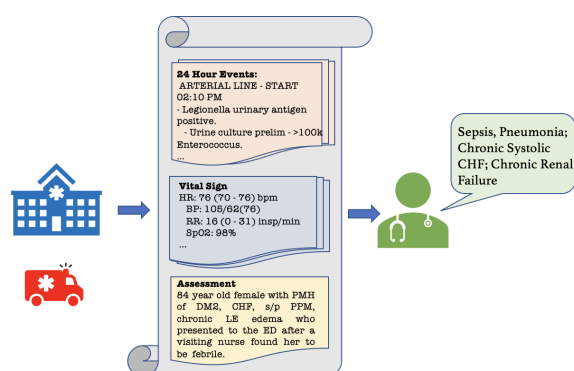

Figure 1: When a sick patient arrives to the hospital, diagnostic evaluations are performed to assess the patient's condition and deduce the problems causing the illness.

progress note for a patient. Healthcare providers write them to document a patient's daily progress and care plan (Brown et al., 2014). The progress note contains both subjective and objective information gathered by the care team, and it is updated daily and serves as the most viewed clinical document by providers. The complexity of the progress note increases as the patient's illness worsens with progress notes collected in the intensive care unit (ICU) representing the sickest patients in the hospital. In the ICU, information and cognitive overload occur frequently, with more opportunities for missed diagnoses and medical errors (Furlow, 2020; Hultman et al., 2019). Automatically generating a set of diagnoses/problems in a progress note may assist providers in overcoming cognitive biases and heuristics and apply evidence-based medicine via information synthesis to accurately understand a patient's condition. These processes may ultimately reduce the effort in document review and augment care during a time-sensitive hospital event (Devarakonda et al., 2017).

Clinical note summarization using natural language processing (NLP) has demonstrated promise in previous work. Hirsch et al. (2015) introduced HARVEST, an EHR summarizer that is currently deployed at point-of-care in a New York hospital. The NLP components of HARVEST include a Markov chain named-entity tagger that identifies diseases explicitly mentioned in clinical notes and a TF-IDF scorer that weighs the importance of the mentions (Lipsky-Gorman and Elhadad, 2011; Hirsch et al., 2015). With the advances of neural methods, recent work has focused on radiology report summarization (Zhang et al., 2018; MacAvaney et al., 2019; Gharebagh et al., 2020) with pointer generator network (See et al., 2017), and doctor-patient conversation summarization (Yim and Yetisgen-Yildiz, 2021; Zhang et al., 2021) with transformer architectures (Vaswani et al., 2017; Raffel et al., 2020). Few investigations apply transformers to *Problem Summarization* progress notes to identify and generate the top diagnoses during a patient's hospitalization.

Problem summarization requires complex cognitive processes to arrive at an accurate diagnosis. When a patient is admitted to the hospital, medical evaluations and diagnostics are initially performed to understand a patient's condition. The review is accompanied by documentation in the progress notes to include pertinent details about the patient's symptoms, medications, physical exam findings, radiology findings, laboratory results, etc. These data are organized in the progress note and used with the physician's medical knowledge to arrive at an assessment of the current problems followed by a treatment plan. The system of nonanalytic and analytic reasoning strategies represent **clinical diagnostic reasoning**, a process involving clinical evidence acquisition with integration and abstraction over medical knowledge to synthesize a conclusion in the form of a diagnosis(Barrows et al., 1980; Bowen, 2006). We hypothesize that to summarize a patient's problems and ultimately develop computerized diagnostic decision support systems, the ability of clinical diagnostic reasoning is the key for NLP systems, a gap in existing NLP literature. In this work, we propose a new summarization task designed to meet the real-world need in the hospital setting as the first step to developing NLP models for clinical diagnostic reasoning. The task is built on a new annotation subset of MIMIC-III (Johnson et al., 2016), a large and publicly available

EHR. We formulate the task as a problem list summarization, as we see the task as the first step in a bigger vision of generating entire notes or sections of notes. Ultimately, the task is designed with our clinical informatics partners to move toward a future real-world application, where a system generating relevant diagnoses can assist healthcare providers and overcome the cognitive burden and information overload. Our contributions include:

- The first knowledge-intensive summarization task towards building NLP systems for computerized diagnostic decision support (sec §3), with an annotated set of clinical notes that are publicly available (sec §4);

- An evaluation of two transformer models for this task, T5 (Raffel et al., 2020) and BART (Lewis et al., 2020), to examine progress in using the state-of-the-art models over a rule-based medical concept extractor (sec §5);

- Domain adaptive pre-training to establish benchmark performance for this task across multiple evaluation metrics (sec §6), with discussion of key challenges and future directions (§7).

2 Related Work

In this section, we provide a brief overview of recently published papers on clinical summarization that use neural methods.

Task setup The stream of recent work on clinical summarization may be divided into two groups: extractive summarization and abstractive summarization. The data corpora are heterogeneous, with multiple note types represented. For extractive sumarization, Liang et al. (2019) proposes a summarization task that extracts sentences from progress notes. Adams et al.(2021) introduces a clinical note summarization task to generate a discharge summary generated from prior notes during hospitalization. More efforts have been made toward abstractive summarization. Several work focus on summarizing radiology reports into an impression, a short piece of text stating the findings from the source image (Zhang et al., 2018; MacAvaney et al., 2019; Gharebagh et al., 2020). Another task is doctor-patient conversation summarization where the output is a summary describing the patient's visit: (Yim and Yetisgen-Yildiz, 2021; Manas et al.,

(*Assessment*) Pt is a 78 y.o female with h.o COPD, HTN, recent MVA with R.ankle/foot fx who presents with hypoxia and LLL infiltrate.
Chief Complaint: Pt does not feel better than at admission, still very fatigued and weak. SOB unchanged. No chest pain. No other complaints.
Allergies: No Known Drug Allergies
Review of systems is unchanged from admission except as noted below Review of systems:

Figure 2: An input example of assessment and subjective sections available in the notes: Chief Complaint, Allergies, Review of systems.

2021; Zhang et al., 2021); or generating clinical notes using both extractive and abstractive summarization: (Krishna et al., 2021). Our work is similar to (Liang et al., 2019) in the emphasis on summarizing problems from progress notes. Yet, Liang et al. (2019) uses a disease-specific dataset (hypotension and diabetes), and formulates the problem as extractive summarization. Our annotations span a broad range of diagnoses across multiple disciplines (surgery, medicine, neurology, cardiology, trauma, etc.) and investigate extractive and abstractive approaches in the task.

Evaluation Prior work has relied on ROUGE (Lin, 2004) as the primary evaluation metric for summarization. Most papers also report human evaluation with aspects of clinical relevancy, factual accuracy and readability (MacAvaney et al., 2019; Gharebagh et al., 2020; Krishna et al., 2021; Yim and Yetisgen-Yildiz, 2021; Zhang et al., 2021). Few have evaluated using a concept F-score, measuring if the predicted summaries contain accurate medical concepts (Liang et al., 2019; Zhang et al., 2021). Our work follows prior work and uses ROUGE, concept F-score, and human evaluation to assess the quality of generated summaries. We also evaluate content quality based on semantic representation using BERTScore (Zhang* et al., 2020) and cosine similarity for sentence embedding.

3 Task Description

Many clinical NLP applications aim to improve physicians' efficiency and decision-making by automatically highlighting essential information from the large body of textual data in the EHR. The goal of Problem Summarization is to identify and generate the problems and diagnoses for the patient's ICU stay. The Problem Summarization task could be developed using a multi-document approach with all notes captured during a hospital encounter. A patient encounter may generate multiple clinical notes (e.g. admission note, transfer note, daily progress notes, etc.), involving different modalities of data such as structured EHR data and radiology images. However, we are particularly interested in facilitating NLP model development for clinical diagnostic reasoning. We define the task as single-document summarization and focus only on a cross-sectional point in time with a single progress note. Our work will show that summarizing a patient's problems over a single progress note is a challenging task and a necessary foundation that requires clinical text understanding and reasoning over sequences of medical concepts.

The progress note is organized in the ubiquitous SOAP format with four components: *S*ubjective, *O*bjective, *A*ssessment, and *P*lan, a documentation method designed to present patient's problems in a highly structured way and developed by Larry Weed, MD (Weed, 1964). Each component has multiple sections gathering patients' information, helping the healthcare providers quickly recognize medical events and active problems. SUBJECTIVE sections are written in natural language and record information about health concerns expressed by patients (e.g. CHIEF COMPLAINTS), and past medical events and history (e.g. ALLERGIES, FAMILY HISTORY). OBJECTIVE sections are primarily structured data, including vital signs, lab tests, medications. ASSESSMENT is a brief description of passive and active diagnoses. It states why the patient is admitted to the hospital and the active problem for the day, usually accompanied by the patient's comorbidities. PLAN section includes multiple subsections, each listing a medical problem and treatment plan. The progress note is time-sensitive EHR data because it is documented daily. As a patient's condition changes and the length of stay increases, the progress note may also increase in length. Another reason for the increasing size is from copy-and-paste behaviour, also known as "note bloat" adding redundant information or noise and hindering the efficiency in data synthesis, which increases the risk for medical error (Rule et al., 2021; Tsou et al., 2017; Shoolin et al., 2013). This reiterates our motivation to develop an NLP system that automatically generates problems and diagnoses to assist providers in clinical workflow and improve diagnostic accuracy.

Our task took Subjective and Assessment sections in progress notes as input and omitted the

Objective sections. Both the Subjective sections and the Assessment section contained information about the reason for admission; therefore, they became the source text (see Figure 2 for an example). The reference summary is a list of problems mentioned in each Plan subsection relevant to the reasons for hospitalization. We will explain the annotation process in the next section. [1]

4 Data

All progress notes were sourced from MIMIC-III, a publicly available dataset of de-identified EHR data from approximately 60,000 hospital ICU admissions at Beth Isreal Deaconess Medical Center in Boston, Massachusetts. We randomly sampled a subset of 768 progress notes and annotated the text spans for the SOAP components. The goal of the annotation was to obtain lists of problems from the Plan subsections. For each Plan subsection, the annotators marked the text span for the Problem, separating the diagnosis/problems from the treatment or action plans. The annotators subsequently determined if the problem was a primary diagnosis (*Direct*), or a past medical problem or consequence from the primary diagnosis (*Indirect*). Two more labels were available for annotating the Plan subsection: *Neither* if the problem or diagnosis was not mentioned in the progress note; *Not Relevant* if the Plan subsection contained non-diagnostic comments such as describing nutrition, prophylaxis, or disposition. Finally, we concatenated the Direct and Indirect problems using semi-colons and used them as *reference summary*. Two medical school students were trained as annotators under the supervision of two board-certified critical care ICU physicians. On the four labels, they achieved a Cohen's Kappa of 0.74 on 10 randomly sampled notes, considered as good quality given the complexity of the task. More details may be found in the Appendix A. [2]

Figure 3 illustrates the task setup. The Direct and Indirect problems were labeled from each Plan subsection using information presented in the input Assessment (entire progress note was also available to the annotators for more information), forming the reference summary (All Problems in the bottom). A total of 1404 and 1599 text spans were labeled as Direct and Indirect Problems, respectively.

The majority of the Direct problems were found in the input Assessment but many of the Indirect problems were not explicitly mentioned in the input Assessment and may be found in other parts of the progress note (*abdominal pain* finding in Subjective or *pneumothorax* finding in chest imaging result of Objective). We also performed medical concept mapping through UMLS (see §5) on the input Assessment and kept the overlap with the reference summaries and categorized them *Explicit Mention of Problems* as an automated labeling approach and baseline. Therefore, the problems represent extractive and abstractive medical concepts. We presented the results across these subgroups assuming the complexity increases as we move from Explicit to Direct to Indirect problems.

5 Experiment Setting

The Unified Medical Language System (UMLS) from the National Library of Medicine is the largest resource containing biomedical concepts and their relationships (Bodenreider, 2004). We applied the concept extractor from QuickUMLS (Soldaini and Goharian), a fast and lightweight Python package, to identify all the medical concepts in the text as the baseline system. Two state-of-the-art seq2seq transformers were selected to compare with the rule-based method: T5 (Raffel et al., 2020) and BART (Lewis et al., 2020). The transformer models are known as data hungry and pre-trained on domain general text, yet our training data was limited in size but full of medical terms. To help the model learn the medical vocabulary and knowledge, we used data augmentation to generate more training samples for our experiments (§5.1 and §5.2).

5.1 Data augmentation

Figure 4 presents a workflow of the data augmentation method across the following three steps: (1) concept identification; (2) synonym mapping; and (3) augmented sample generation. Given an input text, the step of concept identification extracted ngram terms that were matched concepts with UMLS entities, from QuickUMLS. This step was done through a text matcher algorithm using a cosine similarity threshold, setting as Jaccard score with cosine similarity as 1 in our use case. The results returned Concept Unique Identifiers (CUI), a symbolic ID for the medical concept from UMLS. An example output of this step is illustrated in Figure 4: a dictionary of the matched ngrams, e.g.

Input Assessment Assessment and Plan HYPOTENSION (NOT SHOCK) .H/O CIRRHOSIS OF LIVER , ALCO-HOLIC SUPRAVENTRICULAR TACHYCARDIA (SVT) 53 yo female with a history of ETOH cirrhosis c/b persistent leukocytosis and elevated TBili, hx ascites , grade 1 varix , no hx SBP or encephalopathy , recently discharged from), electively admitted for an EUS to evaluate a GE junction polyp (aborted given hypotension), now transferred to the liver service for acute renal failure and?decompensated cirrhosis .

(Plan Subsection 1) Abdominal pain: Per patient, unchanged from previous. S/p Para and KUB, no evidence of SBP or free air. LFTs stable/slightly trending down. Likely [**3-10**] liver disease, though unclear if new or different from past. - Tylenol, Percocet and Ultram PRN pain - Zofran PRN nausea - Trend LFTs
(Plan Subsection 2) Acute renal failure: On admission up to 1.0, peaked at 1.6 and now back to baseline 0.9 after 2 units FFP yesterday. Was likely [**3-10**] ATN. . . .
. . .

Explicit Mention of Problems
hypotension, svt, cirrhosis, leukocytosis, acute renal failure
Direct Problems
Acute renal failure; ETOH cirrhosis
Indirect Problems
SVT with hypotension: This occured in setting of having RA/RV manipulated with wire; Abdominal pain; Leukocytosis; Anemia
All Problems (Reference Summary)
Acute renal failure; ETOH cirrhosis; SVT with hypotension: This occured in setting of having RA/RV manipulated with wire; Abdominal pain; Leukocytosis; Anemia

Figure 3: **Top**: An example assessment input with all the concepts (highlighted in color box) identified through QuickUMLS, a state-of-the-art off-the-shelf medical concept extractor. **Middle**: Two example plan subsections with the the annotated problems, with relation labels omitted. **Bottom**: The reference summary (All Problems) consists of problems annotated as the main reasons for hospitalization (Direct Problems) and secondary concerns (Indirect Problems); explicit mention of the problems is detected by overlapping the concepts identified through UMLS in input and reference summary.

Set	Fine-tuning	Data Augmt.	DAPT
#Notes	700	132k	293k
Input Lens$_{(\sigma)}$	$43.33_{24.75}$	$212.74_{78.39}$	$46.19_{70.95}$

Table 1: Size and average input length (and standard deviation σ) of training set for different experiment settings: the original annotated set for fine-tuning, the data generated from data augmentation method, and DAPT.
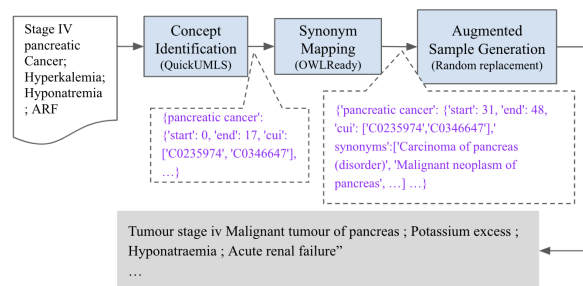

Figure 4: Workflow of the data augmentation method with an input reference summary and output augmented sample

"pancreatic cancer", with start and end character positions and CUIs, e.g. [C0235974]. The mapping module in step 2 found synonyms through CUIs. Here, we used OWLReady (Lamy, 2017) that automatically constructed an UMLS ontology graph, linking the concepts with relations and enabling a quick synonym lookup given a CUI. The synonyms were then passed to the last module for augmented sample generation. The last module randomly chose the synonyms and replaced concepts. An input text may contain n concepts, with each concept having r maximum number of synonyms, the number of combinations of synonyms r^n grows exponentially as n increases. Considering the efficiency, we limited the number of combinations generated by concept replacement to 1000. We ran the pipeline on both reference summary and input assessment, and obtained approximately 132,000 pairs of samples for additional training data. We conducted quality measurement on the augmented samples and report the results in the Appendix B.

5.2 Domain adaptive pretraining with random concept masking

The summarization task requires clinical text understanding and medical knowledge, exposing challenges for models pre-trained from the general domain. Previous work proposed strategies of continuously training the pre-trained language model on domain-specific tasks to enable domain knowledge learning. (Gupta et al., 2021; Pruksachatkun et al., 2020; Gururangan et al., 2020). We followed a similar approach to investigate the effect of domain adaptive pre-training (DAPT) on our summarization task. Specifically, we continuously trained T5 on ASSESSMENT AND PLAN sections from all progress notes in MIMIC, excluding the set of notes for the test set. The result set had 293,000 notes, with the top 3 most frequent note types as Nursing Progress Notes (181k), Physician Resident Progress Notes (61k), and Intensivist Notes (25k).

T5 was trained by random token masking: given a text string, it randomly replaced the token spans with a special tag "*<extra_id_>*" and learned to generate the masked tokens. However, not all words were equally important in our task and we wanted the model to learn clinical semantic types such as symptoms and diseases. Previous work proposed masking on biomedical entities and time expression, achieving performance gains when compared to BERT without entity masking (Lin et al., 2021; Pergola et al., 2021). Inspired by these work, we adapted a concept masking policy where we randomly masked the concepts identified through UMLS. We set the mask token ratio to 15%. For example, the highlighted text in Figure 3 was randomly replaced with the special tag. The statistics of the training set are shown in Table 1.

5.3 Evaluation

We use ROUGE-L (Lin, 2004), a conventional metric in summarization evaluation that based on n-gram overlap, as well as BERTScore (Zhang* et al., 2020), reporting maximum pairwise cosine similarity on word embedding from reference summary and predicted summary. ROUGE fails to recognize synonyms and abbreviations, which are common in biomedical text: e.g., *heart attack* is the same clinical diagnosis as *myocardial infarction*, and *MI* is the abbreviation of *myocardial infarction*. BERTScore compensates this limitation by using contextualized word embeddings from Sap-BERT (Liu et al., 2021), a state-of-the-art BERT encoder (Devlin et al., 2019) for biomedical entity representation that assigns high cosine similarity for synonyms and abbreviations based on UMLS. The reliability of both metrics are validated in literature, thus we report them as main results.

Meanwhile, to better understand the system output, we provide two additional metrics that measure the quality of higher-level information and medical concepts. We took the hidden states of the last layer from SapBERT when taking reference and predicted summary as input, and measure the cosine similarity on sequence embedding (Sent.θ). To evaluate the model's performance in predicting medical concepts, we ran QuickUMLS to get all CUIs from the reference and predicted summaries and computed the F-score. This metric has its own limitation due to the tricky parameter tuning in matching algorithms, causing superfluous or deficient extraction. Regardless, we include them as

approximate solutions towards knowledge-based evaluation for clinical summarization, and leave the metric development for future work.

In the experiments, we set the maximum input and output length to 512 and 128 tokens, respectively. The input text was truncated if the maximum length was exceeded. All experiments occurred on two NVIDIA Tesla V100 32GB GPUs. We used early stopping on the development set during training and saved the models with the highest validation ROUGE-L F-score for evaluation. More implementation details are presented in the Appendix C.

6 Results and Analysis

We evaluated all systems on a test set of 92 progress notes and summaries. Recall that the progress notes contained the Subjective, Objective, Assessment and Plan sections. We set two types of input to the models: (1) Assessment section only (ASSMT), (2) Assessment and Subjective sections (length permitting) (A+SUBJ). Both input settings also had augmented samples from the data augmentation method introduced earlier. We started with a simple rule-based system that was a UMLS concept extractor on the Assessment section. The evaluation metrics across the rule-based, fine-tuned T5 and BART (§6.1), and T5 with domain adaptation pre-training (DAPT, §6.2) are shown in Tables 2 and 3. T5 with DAPT outperformed all other systems and established a benchmark performance for the task. We include a qualitative analysis to provide data-driven insights of the task (§6.3).

6.1 Overall performance of fine-tuned models

Table 2 represents ROUGE-L F-score, cosine similarity on sentence embedding (*Sent.*θ), BERTScore and CUI F-score. Overall, scores dropped from Explicit Mentions to Direct Problems to Indirect Problems, likely due to increasing complexity with more abstractive concepts over extractive concepts. Explicit Mention summarization was the easiest and Indirect Problem summarization was the hardest. The rule-based system outperformed all T5 and BART variants on the Explicit Mentions, given that it identified the obvious entity mentions. For T5 and BART, fine-tuning with augmented samples slightly improved the ROUGE scores. Adding subjective sections (A+SUBJ) did not bring benefits, possibly because most subjective sections are empty in ICU progress note. T5 had more variants with better scores than BART. In our manual in-

Model	Setting	Explicit Mentions				Direct Problems				Indirect Problems				All Problems			
		RL-F	Sent.θ	BS	CUI	RL-F	Sent.θ	BS	CUI	RL-F	Sent.θ	BS	CUI	RL-F	Sent.θ	BS	CUI
Rule-based	ASSMT	**34.45**	58.81	**59.80**	38.97	12.31	**55.33**	40.13	34.23	9.49	**55.58**	**44.46**	33.16	13.45	**68.61**	50.32	43.93
T5	ASSMT	32.77	59.57	57.75	41.73	13.68	53.44	39.72	**36.10**	10.40	54.76	44.16	35.08	14.82	67.49	49.89	44.51
	++	31.76	58.74	57.12	**42.19**	**13.78**	53.65	**40.30**	35.84	**10.55**	54.10	43.48	**35.20**	**15.00**	67.32	50.36	**44.55**
	A+SUBJ	20.24	50.04	47.55	33.44	9.52	51.91	39.72	30.43	7.10	54.14	43.87	30.29	10.89	64.63	49.75	39.02
	++	20.72	**59.64**	57.97	33.56	9.46	53.55	39.52	18.76	7.35	54.69	44.36	14.40	10.93	67.19	**50.42**	24.83
BART	ASSMT	25.70	54.98	52.99	32.49	10.00	53.66	39.08	29.41	8.04	54.66	43.12	29.04	11.56	66.86	48.48	38.36
	++	28.22	57.04	55.16	32.28	10.33	53.40	39.21	30.75	8.29	54.48	44.01	32.08	11.65	66.67	49.23	40.69
	A+SUBJ	18.80	49.19	46.77	26.96	7.04	51.70	38.24	25.30	6.00	54.29	43.71	26.01	9.25	64.95	48.19	34.02
	++	20.23	57.91	54.68	32.91	7.88	53.85	40.21	30.09	6.85	54.61	43.15	30.12	9.84	67.00	49.70	38.72

Table 2: ROUGE-L F-score (RL-F), sentence embedding cosine similarity (*Sent.θ*), BERTScore (*BS*), and evaluation using CUI F-score (*CUI*) from fine-tuning T5 and BART on the two input settings: Assessment (ASSMT), Assessment with Subjective sections(A+SUBJ.) ++ represents the training with data augmentation.

Setting	Model	Token Masking				Concept Masking			
		RL-F	Sent.θ	BS	CUI	RL-F	Sent.θ	BS	CUI
Explicit	ASSMT	32.66	61.34$_{(\uparrow 2.53)}$	56.68	47.10$_{(\uparrow 8.13)}$	29.86	55.87	53.91	40.27$_{(\uparrow 2.14)}$
	++	26.94	59.40$_{(\uparrow 0.59)}$	55.05	42.73$_{(\uparrow 3.76)}$	32.82	58.21	56.80	43.16$_{(\uparrow 4.19)}$
Direct	ASSMT	12.69	53.63	42.40$_{(\uparrow 2.27)}$	35.39$_{(\uparrow 1.16)}$	14.90$_{(\uparrow 2.59)}$	55.48$_{(\uparrow 0.15)}$	47.10$_{(\uparrow 6.97)}$	35.29$_{(\uparrow 1.06)}$
	++	10.44	53.47	43.46$_{(\uparrow 3.33)}$	37.45$_{(\uparrow 3.22)}$	15.76$_{(\uparrow 5.22)}$	56.82$_{(\uparrow 1.49)}$	48.72$_{(\uparrow 8.72)}$	37.74$_{(\uparrow 3.51)}$
Indirect	ASSMT	10.07$_{(\uparrow 0.58)}$	52.72	41.47	38.19$_{(\uparrow 5.03)}$	13.58$_{(\uparrow 4.36)}$	53.44	44.91$_{(\uparrow 0.45)}$	33.56$_{(\uparrow 0.40)}$
	++	8.04	51.84	40.45	37.53$_{(\uparrow 4.37)}$	13.28$_{(\uparrow 4.06)}$	55.02	45.51$_{(\uparrow 1.05)}$	35.10$_{(\uparrow 1.94)}$
All	ASSMT	14.49$_{(\uparrow 1.04)}$	62.40	49.62	40.44	18.72$_{(\uparrow 5.27)}$	64.69	54.03$_{(\uparrow 3.71)}$	42.69
	++	12.12	63.08	50.20	45.58$_{(\uparrow 1.65)}$	18.80$_{(\uparrow 5.35)}$	66.08	55.29$_{(\uparrow 4.86)}$	44.56$_{(\uparrow 0.63)}$

Table 3: Performance of T5 with domain adaptation pre-training using Assessment (ASSMT) as input, under two mask policies: Token Masking and Concept Masking. We report Rouge-L F-score (RL-F), and BERTScore (BS), as well as Sentence embedding cosine similarity (Sent θ) and CUI F-score. Numbers with green background address the highest performance across all results, with subscript number ($\uparrow$) denoting the improvements over rule-based results.
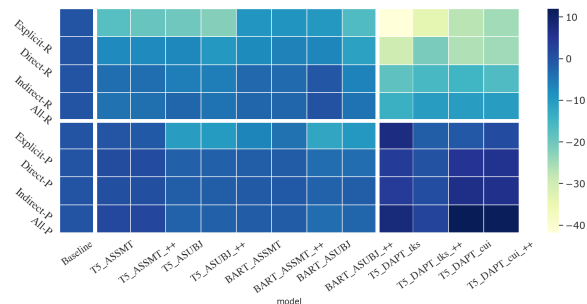

Figure 5: Performance drops (lighter color) and gains (darker color) over baseline (first column) on ROUGE-L Recall (top 4 rows) and Precision (bottom 4 rows). The darker the cell color is, the higher performance gain the model obtains over baseline.

CUI). To highlight the differences before and after DAPT, we showed four scores as well as the performance gained over the baseline system on ASSMT input. Overall, both DAPT settings delivered better performance. The performance gain of T5-DAPT-TKS was mainly from the CUI F-score (+1.16 to +8.13). Superior results were seen from T5-DAPT-CUI, achieving best performance on all setting except for Explicit Mention, yielding large performance gained on ROUGE F score(+2.59 to +5.35) and BERTScore (+0.45 to +8.72).

In addition, Figure 5 includes the ROUGE Recall and Precision drops and gains from all models over baseline. ROUGE Recall measures the content coverage and Precision computes content relevancy in predicted summary (Lin, 2004). All models reported lower recall compared to baseline, indicating their coverage was limited. T5 DAPT variants showed higher gains on precision, yielding the largest gain (+5 to +12 for T5-DAPT-CUI). These results indicate that continuously training T5 with domain vocabulary is a promising direction to solve the task.

vestigation, we find that BART generated text that are not relevant to medical domain[3]. In sum, all fine-tuned model performance were tied with the baseline, which is impressive given that the baseline uses domain knowledge (medical concept).

6.2 The effect of domain adaptation pre-training

Table 3 contains results from training T5 with DAPT and fine-tuned on the annotated set across two methods for masking: random token (T5-DAPT-TKS) and concept masking (T5-DAPT-

---

[3]see Appendix D for example output for all models

---

[3]Semicolons are removed during fine-tuning and evaluation. We manually inserted them back for presentation purpose.

(1)	**[Input Assmt]** 81 yo man with CAD s/p CABG, recent NSTEMI at OSH, HTN, CKD here for ERCP for cholangitis and admitted to ICU with hypotension. **[Reference]** Cholangitis; CAD; HTN; AoCRF **[Prediction]** cad; htn; cholangitis; hypotension; heparin; cholecystitis; dvt
(2)	**[Input Assmt]** 81yo F with h/o ESRD on HD, HTN, HL, acute CHF exacerbation 1 month ago, admitted from OSH with NSTEMI now s/p cath with DES in LAD and LMCA. **[Reference]** SP NSTEMI; HTN; Chronic renal failure; HL; **[Prediction]** esrd; htn; cad; admitted senna chf exacerbation;

Figure 6: Two cherry-picked examples from T5-DAPT-CUI output, with cyan fonts highlighted the correct diseases.[4]

6.3 Qualitative analysis

Besides the numeric metrics reported above, we provide example predicted summaries and qualitative analysis done by a domain expert (a critical care ICU physician). We cherry-picked two examples from T5-DAPT-CUI that best represent the characteristics of medical diagnostic consistency in clinical diagnostic reasoning, and present them in Figure 6. Example 6.1 shows the model performed extractive summarization: it generated both hypertension and hypotension as relevant diagnoses that represent an Indirect label for past medical history of hypertension, and Direct label for an active problem during the hospitalization with hypotension. In example 6.2, the model performed abstraction summarization. The last half of the Assessment highlights a type of heart attack (e.g., "*NSTEMI*") requiring an emergent medical procedure (e.g., "*cath wtih DES in LAD and LMCA*"), and the model summarized a rather complex statement into a single, accurate diagnosis of Coronary Artery Disease in its abbreviated form as "*CAD*".

7 Discussion

Our work begins with a single note in cross-sectional design to build our models; however, a patient's hospitalization is a multi-document workflow with repeated measures of progress notes and other note types across several days and multiple providers. In addition, providers generate their diagnoses via a reasoning process that includes structured data from vital signs, laboratory results, etc. Images and radiology reports are another modality that highlights the multi-modality approach in diagnostic reasoning. Nonetheless, our work opens the door for future research on knowledge intensive clinical summarization. This section includes a discussion of future directions in solving this task.

(1)	a.	**[REF]** Coronary Artery Disease; v fibrillation
	b.	**[PRED]** LVEF-34% - LVEF:36%
		*LVEF (left ventricular ejection fraction) value is abnormally low and indicates a heart disease.
(2)	a.	**[REF]** acute respiratory distress; CAD; chronic kidney injury, Acute Diastolic heart failure
	b.	**[PRED]** ABG on admission 7.25/72/80/32/34
		*ABG (Arterial Blood Gas) reports the abnormally low pH of 7.25 and the abnormally high pCO_2 of 72, and could be from acute respiratory distress and/or acute diastolic heart failure.

Figure 7: Two example reference (**REF**) and predicted summaries (**PRED**) from T5-ALL (input with objective sections).

Exploring structured data The objective section of the progress note contain embedded structured data, delivering rich information regarding patient's problem. Recall the example in Figure 3, the reference summary contains diagnosis: "*Leukocytosis*" (high white blood cell count), "*anemia*"(low red blood cell count). These diagnosis are usually found in laboratory results. To investigate the use of objective sections and structured data, we append both Subjective and Objective sections in chronological order to the Assessment and input to T5 for fine-tuning and evaluation (T5-ALL), and we let the T5 tokenizer truncate the text when it exceeds the 512 token limit. On test set, the scores are too low to report. Yet, we observed that T5-ALL, instead of generating medical concepts, often extracts lines of lab values that strongly associate with the disease in reference summary (see Figure 7.1 and 7.2). This preliminary result indicates the future direction of understanding the association between disease and lab values in summarization.

Incorporating knowledge into models We propose a knowledge intensive summarization task that requires clinical text understanding, knowledge representation and diagnostic reasoning. The experiment results showed that the models pre-trained on medical concepts effectively improved the performance, while challenges remain in understanding the associations among medications, symptoms and disease. Recent work on event extraction and clinical relation extraction incorporates biomedical knowledge graph into pre-trained language models (Huang et al., 2020; Roy and Pan, 2021). Our future work will investigate the incorporation of knowledge graph into seq2seq pre-trained models.

Evidence-based evaluation Medical diagnosis is a critical component of effective healthcare but misdiagnosis is a major contributor to medical errors, especially in critical care settings where quick decision-making is needed. Medical diagnoses predicted by systems that are not redundant must

be contextually relevant to the data gathered in a progress note to achieve valid reasoning. We believe an automated evaluation method for problem summarization should assess the knowledge representation, non-redundancy, and evidence relevancy, and the automated metrics used in our work cover partial aspects. Recently, Moramarco et al. (2021) studied a fact-based evaluation for medical summarization using human evaluation, which we plan to carry out in future work.

8 Conclusion

We propose a problem summarization task that address diagnostic reasoning, and show that T5 with DAPT achieves benchmarking performance for the task, but some key challenges remained. Our work lays the ground for future research on knowledge fused clinical summarizers as well as real-world clinical diagnostic decision support system. Future work will investigate the uses of structured data, evidence-based evaluation metric and better models for knowledge representation and summarization.

Ethical Statement The use of the data in this research came from a fully de-identified dataset (contains no protected health information) that we received permission for use under a PhysioNet Credentialed Health Data Use Agreement (v1.5.0). The study was determined to be exempt from human subjects research. All experiments followed the PhysioNet Credentialed Health Data License Agreement.

Medical charting by providers in the electronic health record is at-risk for multiple types of bias. Our research focused on building a system to overcome the cognitive biases in medical decision-making by providers. However, statistical and social biases need to be addressed before integrating our work into any clinical decision support system for clinical trials or healthcare delivery. In particular, implicit bias towards vulnerable populations and stigmatizing language in certain medical conditions like substance use disorders are genuine concerns that can transfer into language model training (Thompson et al., 2021; Saitz et al., 2021; Karnik et al., 2020). Therefore, it should be assumed that our corpus of notes for this task will carry social bias features that can affect fairness and equity during model training. Before the deployment of any pre-trained language model, it is the responsibility of the scientists and health system to audit the model for fairness and equity in its performance across disparate health groups (Saleiro et al., 2018). Fairness and equity audits alongside model explanations are needed to ensure an ethical model trustworthy to all stakeholders, especially patients and providers.

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.

Howard S Barrows, Robyn M Tamblyn, et al. 1980. *Problem-based learning: An approach to medical education*, volume 1. Springer Publishing Company.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Judith L Bowen. 2006. Educational strategies to promote clinical diagnostic reasoning. *New England Journal of Medicine*, 355(21):2217–2225.

PJ Brown, JL Marquard, B Amster, M Romoser, J Friderici, S Goff, and D Fisher. 2014. What do physicians read (and ignore) in electronic progress notes? *Applied clinical informatics*, 5(02):430–444.

Murthy V Devarakonda, Neil Mehta, Ching-Huei Tsou, Jennifer J Liang, Amy S Nowacki, and John Eric Jelovsek. 2017. Automated problem list generation and physicians perspective from a pilot study. *International journal of medical informatics*, 105:121–129.

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.

Bryant Furlow. 2020. Information overload and unsustainable workloads in the era of elec-

tronic health records. *The Lancet Respiratory Medicine*, 8(3):243–244.

Sajad Sotudeh Gharebagh, Nazli Goharian, and Ross Filice. 2020. Attend to medical ontologies: Content selection for clinical abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1899–1905.

Yash Gupta, Pawan Sasanka Ammanamanchi, Shikha Bordia, Arjun Manoharan, Deepak Mittal, Ramakanth Pasunuru, Manish Shrivastava, Maneesh Singh, Mohit Bansal, and Preethi Jyothi. 2021. The effect of pretraining on extractive summarization for scientific documents. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 73–82.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Jamie S Hirsch, Jessica S Tanenbaum, Sharon Lipsky Gorman, Connie Liu, Eric Schmitz, Dritan Hashorva, Artem Ervits, David Vawdrey, Marc Sturm, and Noémie Elhadad. 2015. Harvest, a longitudinal patient record summarizer. *Journal of the American Medical Informatics Association*, 22(2):263–274.

Kung-Hsiang Huang, Mu Yang, and Nanyun Peng. 2020. Biomedical event extraction with hierarchical knowledge graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1277–1285.

Gretchen M Hultman, Jenna L Marquard, Elizabeth Lindemann, Elliot Arsoniadis, Serguei Pakhomov, and Genevieve B Melton. 2019. Challenges and opportunities to improve the clinician experience reviewing electronic progress notes. *Applied clinical informatics*, 10(03):446–453.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Liwei 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(1):1–9.

Niranjan S Karnik, Majid Afshar, Matthew M Churpek, and Marcella Nunez-Smith. 2020. Structural disparities in data science: a prolegomenon for the future of machine learning. *The American Journal of Bioethics*, 20(11):35–37.

Kundan Krishna, Sopan Khosla, Jeffrey P Bigham, and Zachary C Lipton. 2021. Generating soap notes from doctor-patient conversations using modular summarization techniques. 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 4958–4972.

Jean-Baptiste Lamy. 2017. Owlready: Ontology-oriented programming in python with automatic classification and high level constructs for biomedical ontologies. *Artificial intelligence in medicine*, 80:11–28.

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.

Jennifer Liang, Ching-Huei Tsou, and Ananya Poddar. 2019. A novel system for extractive clinical note summarization using EHR data. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 46–54, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2021. Entitybert: Entity-centric masking strategy for model pretraining for the clinical domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 191–201.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

S Lipsky-Gorman and N Elhadad. 2011. Clinnote and healthtermfinder: a pipeline for processing clinical notes. *Columbia University*.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238.

Sean MacAvaney, Sajad Sotudeh, Arman Cohan, Nazli Goharian, Ish Talati, and Ross W Filice. 2019. Ontology-aware clinical abstractive summarization. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1013–1016.

Gaur Manas, Vamsi Aribandi, Ugur Kursuncu, Amanuel Alambo, Valerie L Shalin, Krishnaprasad Thirunarayan, Jonathan Beich, Meera Narasimhan, Amit Sheth, et al. 2021. Knowledge-infused abstractive summarization of clinical diagnostic interviews: Framework development study. *JMIR Mental Health*, 8(5):e20865.

Francesco Moramarco, Damir Juric, Aleksandar Savkov, and Ehud Reiter. 2021. Towards objectively evaluating the quality of generated medical summaries. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 56–61.

Gabriele Pergola, Elena Kochkina, Lin Gui, Maria Liakata, and Yulan He. 2021. Boosting low-resource biomedical qa via entity-aware masking strategies. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1977–1985.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Arpita Roy and Shimei Pan. 2021. Incorporating medical knowledge in bert for clinical relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5357–5366.

Adam Rule, Steven Bedrick, Michael F Chiang, and Michelle R Hribar. 2021. Length and redundancy of outpatient progress notes across a decade at an academic medical center. *JAMA Network Open*, 4(7):e2115334–e2115334.

Richard Saitz, Shannon C Miller, David A Fiellin, and Richard N Rosenthal. 2021. Recommended use of terminology in addiction medicine. *Journal of Addiction Medicine*, 15(1):3–7.

Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. 2018. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*.

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.

J Shoolin, L Ozeran, C Hamann, and W Bria Ii. 2013. Association of medical directors of information systems consensus on inpatient electronic health record documentation. *Applied clinical informatics*, 4(02):293–303.

Luca Soldaini and Nazli Goharian. Quickumls: a fast, unsupervised approach for medical concept extraction.

Hale M Thompson, Brihat Sharma, Sameer Bhalla, Randy Boley, Connor McCluskey, Dmitriy Dligach, Matthew M Churpek, Niranjan S Karnik, and Majid Afshar. 2021. Bias and fairness assessment of a natural language processing opioid misuse classifier: detection and mitigation of electronic health record data disadvantages across racial subgroups. *Journal of the American Medical Informatics Association*, 28(11):2393–2403.

Amy Y Tsou, Christoph U Lehmann, Jeremy Michel, Ronni Solomon, Lorraine Possanza, and Tejal Gandhi. 2017. Safe practices for copy and paste in the ehr. *Applied clinical informatics*, 26(01):12–34.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Lawrence L Weed. 1964. Medical records, patient care, and medical education. *Irish Journal of Medical Science (1926-1967)*, 39(6):271–282.

Wen-wai Yim and Meliha Yetisgen-Yildiz. 2021. Towards automating medical scribing: Clinic visit dialogue2note sentence alignment and snippet summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 10–20.

Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R Gormley. 2021. Leveraging pretrained models for automatic summarization of doctor-patient conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3693–3712.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. 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 *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 204–213.

A   Annotator Training

We recruited two medical students as annotators who received training in medical school curriculum in SOAP note documentation. A three-week orientation and training was conducted by one of the critical care physicians. Each annotation achieved an inter-annotator agreement with a kappa score above 0.80 with the adjudicator. Another round

Input	Sent.θ	Jaccard	Length Diff.
ASSMT	89.00	37.85	6.13 (4.12)
SUMM	83.14	14.43	9.42 (5.99)

Table 4: Quality measurement on augmented input assessment (ASSMT) and reference summary (SUMM). For ever pair of original and augmented sample, we report cosine similarity between text embedding (θ), Jaccard token overlap, and mean and standard deviation (σ) of length difference.

Hyper-parameter	Setting
Optimizer	AdamW
Epoch	10 (with early stopping)
Learning rate	1e-3, 1e-4
Batch size	256
Gradient accumulation	True

Table 5: Hyperparameters for T5 DAPT

of training was performed on 200 notes and the inter-annotator agreement was measured between annotators and the adjudicator. The annotation was reviewed if the kappa score is below 0.80 threshold.

B   Quality Measure for Data Augmentation

The quality of augmented data directly affected the training process. To ensure a high-quality training corpus, we randomly selected 2,000 pairs of augmented samples. We evaluated how well the meanings were preserved in the augmented sample, and how much lexical variance was introduced into the augmented samples. We reported cosine similarity between the embedding pairs for quality of meanings, and we reported Jaccard similarity for degree of string overlap. Specifically, given a pair of the original sample and the augmented sample, we generated a text embedding through SapBERT (Liu et al., 2021), a BERT encoder pre-trained to represent biomedical entities using UMLS. We expected a high cosine similarity if the augmented samples expressed the same meanings as the original samples. We ran Jaccard similarity by treating the samples as lists of tokens, and expected a low Jaccard score if there were new terms introduced in the augmented samples, e.g. *ARF* and *Acute Renal Failure*. We also reported the mean and standard deviation of the length differences between original and augmented samples (Table 4). On both input assessment and reference summary, the cosine similarity between original and augmented samples was higher than 0.80. Assessment input contained more words that were not biomedical

Input ALCOHOL ABUSE Assesment: This is a 37 year-old male with a history of alcoholism with multiple admissions for detox and history of DTs/withdrawal seizures who presents with intoxication followed by withdrawal.	
Reference ETOH intoxication; Elevated lipase	

FT-T5 ALCOHOL ABUSE Assesment: This is a 37 year-old male with multiple admissions for detox and history of DTs/withdrawal seizures who presents with intoxication followed by withdrawal.

FT-BART ALCOHOL ABUSE DIAGNOSIS Assesment: This is a 37 year-old male with a history of alcoholism with DTs/withdrawal seizures. He presents with intoxication followed by withdrawal. Alcoholics Anonymous: He is an alcoholic with multiple admissions for alcohol abuse and drug addiction. This male is the only person in the state of Alabama with alcohol dependence.

T5-DAPT-TKS This is a 37 year-old male with history of DTs/withdrawal seizures who presents with intoxication followed by withdrawal

T5-DAPT-CUI history alcoholism history history of dts/withdrawal seizures admissions alcohol abuse history

Figure 8: Given an input assessment, we show the reference summary, example output from fine-tuning T5 and BART, and T5 DAPT with token masking and concept masking. The red fonts show the information that is outside the input text.

Hyper-parameter	Setting
Optimizer	Adam
Epoch	10 (with early stopping)
Learning rate	1e-4, 1e-5, 1e-6
Batch size	4
Task Prefix (t5)	"summarize:"
Encoder max length	512
Decoder max length	128
Beam size	10
Length penalty	1
no repeat ngram size	2

Table 6: Hyperparameters for fine-tuning T5 and BART

concepts; thus, the augmented sample had a greater proportion of overlapping text than the reference summary. Both had more than 6 token differences in length. In conclusion, our proposed strategy of data augmentation successfully produced a high quality training corpus.

C Hyperparameters

Here we report the hyper-parameters we used for T5 DAPT experiments in table 5, and fine-tuning for t5 and BART in table 6. The input length to both T5 and BART is set to 512 tokens. On the training data, the average length of input assessment is 43.33 tokens, and the average length of input and subjective sections is 70.97 tokens. Therefore the maximum encoder length is appropriate for our task.

D Model Example Output

Figure 8 shows the example output from fine-tuning T5 and BART, and T5-DAPT with token masking as well as concept masking policy. T5-DAPT-CUI extracts medical concepts. FT-T5 and T5-DAPT-Tks extract sequence of text from input

assessment. FT-BART produces text with information that is not mentioned in the input (red fonts).