Extracting Clinical Information from Unstructured EHRs using Language Models, and its Role in Disease Prediction

Sitong Zhou

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2026

Reading Committee:

Mari Ostendorf, Chair

Meliha Yetisgen, Chair

Hannaneh Hajishirzi

Fei Xia

Program Authorized to Offer Degree:
Electrical Engineering

University of Washington

Abstract

Extracting Clinical Information from Unstructured EHRs using Language Models, and its Role in Disease Prediction

Sitong Zhou

Co-Chairs of the Supervisory Committee:
Professor Mari Ostendorf
Electrical & Computer Engineering

Professor Meliha Yetisgen
Biomedical Informatics & Medical Education

Clinical unstructured data contain critical information for clinical decision making, such as symptoms and radiology findings, that can complement structured EHRs and often add greater details. However, clinically relevant information can be buried in abundant EHR unstructured notes, which can challenge physicians to review. In addition, large volumes of texts can include irrelevant information to secondary machine learning applications. We aim to develop language model–based information extraction (IE) methods to extract clinically critical information from EHR texts, supporting human review and secondary clinical decision applications. We first develop robust event extraction methods using supervised learning to identify clinical events at the sentence level, and improve their domain generalization across different domain shifts. In one study on symptom event extraction, we demonstrate that two strategies, adaptive pretraining on unstructured EHRs and masking frequent symptoms during training, improve domain generalization when using an encoder-only language model. In a second study on radiological findings extraction, we show that generative LMs generalize better than encoder-only models in categorizing minority classes, and further training them on decomposed, simpler subtasks improves generalization to complex tasks when subtask dependencies are shifted across domains. In addition to event extraction from isolated reports, we present longitudinal summarization of radiology reports as an additional IE task to track

radiological findings and capture temporal changes not reflected in individual reports. We frame longitudinal summarization as a timeline generation task that groups related findings across time, and introduce RadTimeline as an evaluation dataset, and propose an LLM-based approach that achieves good recall of lung findings and human-comparable grouping of gold-standard findings without training data. Finally, we apply information extraction to extract risk factors from longitudinal EHRs for a secondary-use clinical application, a lung cancer prediction task. We create a lung cancer case-control cohort, where each patient has a 5-year longitudinal EHR history and a lung cancer outcome within three years. We find that COPD, smoking status and radiology abnormality information extracted from unstructured notes can complement the structured EHRs, and improve lung cancer risk prediction performance. Using a transformer-based risk prediction model, we further compare different representations of longitudinal risk factors across model variants and input orderings, finding that there is no benefit from including findings from reports beyond a 6-month window.

ACKNOWLEDGMENTS

I am grateful for my two wonderful advisors, Mari Ostendorf and Meliha Yetisgen, for their extensive advising efforts and financial support throughout this long term journey. They give me lots of valuable feedback to develop my work and strengthen my academic skills, and greatly help build my confidence and communication skills that would benefit me far beyond academia. They are also always encouraging and believe in me, and are protective during the pandemic.

I want to thank my committee members, Hannaneh Hajishirzi, Fei Xia, Lucy Wang, for their help and guidance, and being inspiring.

I want to thank for advising from Kevin Lybarger, Nic Dobbins and Matthew Thompson for the advice related to ClinicalNLP and clinical knowledge.

I am fortunate to work within two groups, TIAL and BioNLP, with inspiring labmates working on different fields, I really enjoy the intelectual stimulus and social comfort. I am especially grateful to my labmates, Trang Tran, Ellen Wu, Yushi Hu, Velvin Fu, Namu Park, Bin Han, Sihang Zeng, and Avery Yu, for their invaluable support and thoughtful suggestions for my projects.

I want to thank all my roommates and many neighbors over different neighborhood areas for the comfort. I also want to thank my violin tutor and Seattle symphony volunteer program for boosting my mental energy and clarity.

I also want to thank my longtime friends, Du Xiaohan, Sherry Wu, and Feng Yijing, who always uplift my spirits, and Duan Yuqing and Huang Yuan for making Seattle feel like home.

Finally I want to thank all my family members, for the unconditional love, support, and also the freedom to let me live in a remote country to do what I like.

DEDICATION

To my supportive advisors, family, and friends, who make this journey possible.
To my four grandparents, for giving me a golden childhood. Especially to my grandfather Zhu Wenbi, a creative and compassionate person, and a civil engineer who designs roads, bridges, and highways, who left me the legacy of joyful spirits and fun hobbies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

Electronic health records (EHRs) contain rich clinical information describing patient status, treatment details, and disease progression [6]. EHRs are longitudinal and contain both structured data (e.g., diagnosis codes and laboratory results) and unstructured clinical narratives, such as clinical notes generated from patient–clinician interactions and radiology reports derived from image interpretation [7]. Unstructured textual data can capture facts and details that are missing from structured data and can complement it [8]. For example, symptoms, which are helpful for understanding treatment response, are often richly documented in clinical notes, with critical details such as severity [3]. Radiology reports contain findings that are critical for assessing the risk of many diseases such as lung cancer, and include key details such as lesion size and change over time [9]. However, for clinical practitioners, reading lengthy longitudinal textual records is time-consuming and can cause cognitive overload. For secondary machine learning algorithms for outcome prediction, raw unstructured EHRs include additional information that is irrelevant to the outcomes, and can exceed the maximum length of a model. In this dissertation, we aim to extract clinically critical information from EHR texts using language models, in order to facilitate human review and secondary outcome prediction applications. We develop tools for two types of information extraction (IE) formats, event extraction on isolated reports, and longitudinal summarization on longitudinal reports, then apply them to a lung cancer prediction task.

We first develop robust clinical information extraction methods that extract clinical events at the sentence level, where each event contains a key entity and related details. Though the advanced large language models (LLMs) enable learning IE tasks with zero or a few labels [10, 11], supervised learning methods on adequate task-specific annotations from at least one hundred notes can give the state-of-art for many tasks with smaller language models (LM) that are computationally cost-effective [12, 13, 14]. However, clinical texts can vary

from institutions, note types and patient cohorts, and costly medical expert annotations for model development are often limited to a certain collection of notes sharing some attributes, referred to as a domain. When such models are expected to process out-of-domain notes with different attributes, they are at risk for performance degradation under domain shift. The model might overfit the training data from one domain, learning features that are not transferable, even harmful, when applied to another domain. We aim to build IE tools that work robustly on diverse notes. We analyze two clinical IE cross-domain scenarios where performance degrades under different domain shifts, and propose domain generalization methods to improve encoder-only transformers for symptom extraction and generative LMs for radiological finding extraction. In the study on symptom extraction, we show that two strategies, additional adaptive pretraining on a clinical-domain, encoder-only language model and masking frequent symptoms in the training data, both improve domain generalization. In the other study on radiological findings extraction, we show that generative LMs can generalize better than the encoder-only LM in categorizing minority classes, and training generative LMs on decomposed, simpler tasks can improve generalization on complex tasks, when subtask dependencies are shifted across domains.

The above event extraction models are designed for isolated reports, and may not capture details of changes in clinical information that would be seen in reading longitudinal reports. However, reading longitudinal notes throughout a patient's history can be time-consuming and cause cognitive overload. To reduce this burden, we develop longitudinal summarization methods that group related findings spread across time. While artificial intelligence (AI) tools for summarization are effective for many tasks, most work on multiple documents has involved unstructured summaries, which complicates fact-checking and is not ideal in high-stakes clinical scenarios. A verifiable structured summarization format Verma et al. [15] has been proposed in which each fact is linked to its source document and organized into human-curated topics. However, a pre-defined set of topics is not practical for representing radiological findings for specific patients, where multiple nodules may be observed in the same region of the lung, for example. We build a structured summarization with topics adaptable to patients, and introduce a new two-dimensional timeline format for structured summarization that makes it easy to see temporal changes. We create RadTimeline, an

evaluation dataset with lung-related radiological findings within longitudinal chest-related imaging reports, and propose a timeline generation approach based on large language models (LLMs) that groups related findings using generated interpretable group names. We show this approach has good recall of findings, despite including irrelevant and duplicated findings, and can group gold findings comparable to humans.

Finally, we apply the extracted information from longitudinal EHRs in a secondary-use clinical application, a lung cancer prediction task. Lung cancer is the leading cause of cancer-related deaths in the US, and early detection can reduce mortality. Extracted risk factors from textual EHRs can complement structured data, and include more risk factor details that are related to lung cancer risk. For example, smoking status is sometimes recorded in clinical narratives but not in the structured data. Important details of radiological findings, such as size and change trend, are often described in radiological reports and beyond the granularity of the International Classification of Diseases (ICD) diagnosis coding system. We first create a lung cancer case-control cohort, where each patient has a 5-year longitudinal EHR history and a lung cancer outcome within three years. We find that using LLMs to extract numerical or categorical information from longitudinal unstructured data improves data completeness for risk factors and improves the risk prediction performance over a structured-EHR-only baseline. To represent longitudinal radiological findings, we compare using a sequence of time-ordered lung findings extracted by IE and LLMs, LLM-extracted findings ordered by finding groups, and a standard LLM vanilla summary. That information is fed into an encoder-only LM model for risk prediction with supervised learning. We show that the time-ordered sequences from IE and LLM give the best results, and that the longer information beyond 6 months does not provide additional advantage.

There are three main contributions of this thesis. First, we propose domain generation methods that mitigate domain shift performance degradation for clinical IE models to robustly extract information from clinical sentences. Second, we propose a timeline generation to summarize longitudinal records. Third, we predict lung cancer risk using extracted risk factors, and show improvement over structured record baseline, and compare different representations of longitudinal risk factors.

This thesis is organized as follows. Chapter 2 presents background of EHRs and describe

event extraction and summarization tasks and the associated the challenges of applying language models (LMs). Chapters 3 and 4 describe domain generalization strategies for clinical event extraction under domain shifts using fine-tuned LMs. Chapter 3 focuses on symptom extraction using encoder-only LMs across institutions, while Chapter 4 examines the use of generative LMs for extracting radiological findings across imaging modalities. Chapter 5 introduces a structured summarization task to reveal temporal changes in radiological findings from longitudinal EHRs. Chapter 6 presents a secondary application, a lung cancer prediction task, that leverages the extracted risk factors specific to lung cancer. Finally, Chapter 7 summarizes the major findings and suggests future research directions.

Chapter 2

BACKGROUND

2.1 Motivation

EHRs contain longitudinal clinical records of patients in heterogeneous sources, including unstructured data such as clinical notes and radiological reports, written by clinical practitioners as free-text narratives, as well as structured data collected through caregivers logging with machines, filling forms, or entering codes such as International Classification of Diseases (ICD) diagnosis codes [7]. However, structured data does not always sufficiently reflect patient information. For example, ICD codes are often used for billing purposes, they may be missing, miscoded, or for suspected rather than confirmed diagnoses. Compared to using only structured data, combining clinical narratives improves completeness and accuracy of important information that supports clinical decisions [4, 8, 16, 17, 18, 19], and also captures details beyond the structured data schema [9, 3, 20, 4, 21, 22]. In addition to structured data, unstructured data has shown effectiveness in improving data completeness of information such as social determinants of health (SDOH) [4, 23], diagnoses [18], and symptoms [19, 24]. Combining unstructured EHRs with structured ones improves the accuracy of identifying patients with certain diseases [18]. Unstructured data also contains critical details beyond the structured data schema, such as negation and temporal changes. As figure 2.1 illustrates, a clinical note can include symptom nuances such as negation, change, severity, and duration [3]; smoking-related details such as assertion, amount, duration [4] and quit dates of former smokers [21]. Radiology notes also contain radiological finding details such as lesion size and change trend, which are crucial for risk assessment [9].

Standardizing clinical information into structured data in a certain semantic schema, can help standardize free-text information and facilitate secondary applications for large cohorts [22, 25, 6]. Successful applications include disease prediction [26], phenotyping [18, 20, 16], clinical trial recruitment [17, 21], and cohort analysis [24]. Extracted smoking quantity and

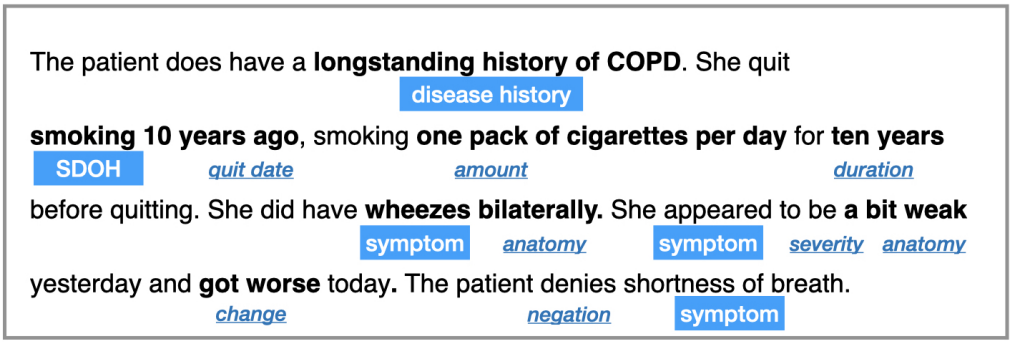

Figure 2.1: A synthetic clinical narrative about progress note.

temporal details can accurately identify patients for low-dose computed tomography (LDCT) scanning, to identify the criteria of "at least 20 pack-years, current smoker or have quit within 15 years" [21]. Symptom events extracted from free-text notes can be merged with coded symptom data to analyze early symptoms of lung cancer prior to diagnosis [24].

EHRs contain abundant notes from different sources and times, and reading all unstructured data is challenging under time-restricted clinical scenarios. Summarization of unstructured EHRs can physicians' reviewing time and cognitive burden by helping identify the relevant clinical information buried within extensive unstructured notes [27] and synthesize information to support clinical decisions [28, 15]. Extractive summarization methods for discharge notes have been developed to synthesize a patient's medical history from diverse data sources, aiming for reducing time and better communication and care, especially helpful when caring for chronically ill patients whose records may contain hundreds of clinical notes [28]. Liang et al. [29] developed a disease-specific extractive summarization system to extract only the most important information relevant to hypertension or diabetes mellitus from a single note. A query-focussed extractive summarization method was developed to help radiologists, who typically have fewer than 10 minutes to complete their interpretation, identify snippets of textual information relevant to a particular potential diagnosis [27]. Verma et al. [15] developed an extract-then-abstract summarization method that prioritizes and synthesizes clinically relevant information from the longitudinal EHRs of patients

admitted with heart failure, and found frequent use of the summaries significantly shortened the physicians' time spent in answering patient-specific clinical questions.

2.2 Clinical Event Extraction

Event extraction can be formulated as span-based IE, which detects entities by classifying enumerated spans and identifies relations by classifying pairs of spans. An event is identified by a trigger and characterized by multiple arguments linked to the trigger, and can be extracted by joint entity and relation extraction methods. Methods using encoder-only transformers [30, 1, 31, 32, 3, 30, 4, 1, 31], mostly based on variations of BERT ([33, 34, 35, 36, 37], achieve success when sufficient training data is available and is similar to the test distribution. DYGIE++ [30] uses graphs to propagate contextual information on top of the transformer encoder. SpERT [1] employs more lightweight entity and relationship classifiers without compromising performance. PURE [31] uses different transformer encoders for entity and relation classification and fuses entity information into relation extraction by inserting markers.

However, the above encoder-only architectures represent entity and relation labels as numbers, therefore the learned representation of the labels depend totally on training data, which challenges situations with domain shifts and imbalanced classes. To infuse task information besides training data, Li et al. [38] formulate the task as multi-hop question answering (QA) and create questions using templates, in which the entity and relation types are described as text. Later, generative approaches [39, 40] propose to frame tasks in a text-to-text format, and describe the task in natural language in order to reduce the reliance on training data, showing promising results in question answering. Schema-based prompts containing task-specific description and ontology are found effective to help generative models with structured outputs, such as semantic parsing [41], dialog state tracking [42, 43, 41], and IE [44]. Besides general domain generative models, more domain-specific versions are developed for biomedical and clinical domains [5, 45, 46, 47, 41].

Larger generative models generalize better with limited training data. GPT-3 [40] shows zero-shot and few-shot ability on clinical IE, requiring minimal expert annotation [10]. However, their increased latency and computation costs during inference hinder scalability

for clinical applications. Efforts for reducing inference costs include switching to a smaller model [48], distilling from larger models [49, 50], and reducing the number of decode passes during inference [31].

Our studies on event extraction focus on moderately-sized finetuned models under 220 million parameters, including both encoder-only and generative models, and take latency into consideration.

2.3 Domain Generalization for Supervised Learning

When trained on adequate in-domain labeled data, transformers can achieve near-human-level performance measured by in-domain evaluation metrics in clinical IE tasks [9, 51, 35, 45, 5]. However, clinical applications are expected to apply to a variety of domains. In some cases, performance is degraded under distribution shifts. For clinical QA tasks, a significant drop in F1 scores is observed when evaluating on unseen clinical notes [52], and BERT-style models are found sensitive to natural distribution shifts across topics [53]. For radiology image classification, performance degradation has been observed when models are applied across institutions. [54, 55]. For mortality prediction using EHR inputs, model performance decays when trained on historical data and evaluated on future data [56].

Prior domain generalization (DG) work can be categorized into four scenarios based on the information available in the target domain. The first case is to generalize the model to an arbitrary unknown domain [57, 58]. The second is to apply the model to a domain which we have some knowledge of the domain but cannot access the data. In the last two senarios, target domain data is available, such cases are often referred to as domain adaptation (DA). The third case is when unlabeled data from the target domain is available for unsupervised domain adaptation (UDA) [59]. The last is when some target domain labeled data exists [60]. Our work focuses on the second and third cases, which represent common clinical scenarios where the target domain is known, often easily identifiable through clinical metadata, with abundant unlabeled target domain data available for optional usage, but no labeled target domain data due to the costs of expert annotations. The key to success is to learn domain-invariant features that are informative to task classifiers. Some methods learn domain-invariant features using domain-adversarial training [61, 62, 63]; some reweight

or select source domain training data similar to target domain data [64, 65, 62, 66, 67]; some manipulate datasets by pseudo-labeling [68, 69] or generating new examples [70, 71]; some create auxiliary tasks [72] such as optimizing LM objectives for target domains [73, 74, 75], solving jigsaw puzzles [76], and predicting coarse entity types [77].

Previous work across different tasks shows that machine learning methods sometimes learn spurious features that are predictive during training but do not generalize. Geirhos et al. [78] describe a shortcut learning phenomena in which models tend to learn tasks with the least amount of effort without sufficient understanding of their underlying principles. For example, in natural language inference (NLI) tasks [79] that determine relations between hypothesis and premise statements, the hypothesis alone shows predictive power. In digit image classification, models rely on background noise [80] that is not the generalizable. Performance degradation can sometimes be attributed to training data containing information that is not essential to the task, such as auxiliary inputs [81] and confounding variables [54].

To avoid the unwanted short-cut learning, simplifying training data by removing unnecessary information to the task can help generalization. Liu et al. [77] predict entity types at a coarse level for IE tasks, and Nestor et al. [56] group input items into expert-defined concepts for mortality prediction. For generative models specifically, the dependency between subtasks of a complex task can also be unwanted. Press et al. [82] report that when answering multi-hop questions difficult to find on the internet, performance is worse than solving sub-problems separately, referred to as compositionality gap. Breaking questions into several reasoning steps mitigates this problem [83, 82] for LLMs. For finetuned models, task decomposition also shows better generalization from training to in-domain test distribution. Decomposing tasks into multiple subtasks benefits multi-hop QA [84], IE [31], and DST [42]. To complete multiple subtasks, small models can perform multiple decoding runs [84], but this may impact latency in real-time applications. However, research is limited in enabling smaller models to complete a sequence of subtasks in a single inference run. Our study proposes data manipulation methods to remove unnecessary information from the inputs, rather than augmenting with synthetic labeled data which can potentially introduce noise [68, 70]. Specifically, we remove domain-specific phrases when training encoder-only LMs, and divide the training data into shorter subtask sequences for generative LMs, enabling

them to complete the full task within a single inference run.

Other than improving labeled data, unlabeled target domain data can be used to pretrain LMs for better transfer learning across tasks [33, 39]. One explanation for transfer learning ability of pretrained models is that they learn latent concepts that are critical to downstream tasks [85]. The pretraining LMs also encode clinical knowledge that can be shared across tasks [86]. Pretraining on data relevant to the task domain is beneficial for scientific [30, 36, 1], biomedical [34], clinical domains [35, 37]. A variety of off-the-shelf domain-specific LMs are available, including SciBERT [36] for scientific articles, BioBERT [34] and PubMedBERT [37] for biomedical literature, and Bio+Clinical BERT [35] for clinical notes. The Bio+Clinical BERT is pre-trained on MIMIC III [87], which consists of clinical notes of patients admitted to intensive care units (ICU) from a single medical center. The off-the-shelf clinical LM may not be adequate for representing our task since our data covers a broader range of patient populations from different medical systems. Gururangan et al. [88] introduce an adaptive pretraining strategy that does further pretraining on close-domain texts. This is suitable for private datasets that are beyond the knowledge of the off-the-shelf models trained on public data. This suggests potential benefits to pretrain the clinical domain LM further on unlabeled texts from institutional EHRs.

Since pretrained models are difficult to update, unlabeled target-domain text can also be used to retrieve relevant target-domain contexts without further training [89]. Integrating models with supplementary contexts has shown benefits in knowledge-intensive tasks [90, 91]. Generative models can utilize knowledge prompts from external knowledge sources [92, 93]. In our work, we retrieve contexts from the unlabeled clinical notes in EHRs instead of from external resources.

Domain shifts in EHRs can offer valuable lessons into real-world data shifts that cannot be replaced by studies conducted under synthetic domain shifts. Despite many domain generalization methods developed under synthetic environments [94], robustness under synthetic data shift does not transfer to real-data shifts [95]. Real clinical distribution shifts are studied on tasks such as image classification [96], mortality prediction [56, 54], and QA [53]. In some realistic domain shift benchmarks [96, 54, 53], no single method is effective for all distribution shift scenarios. Real domain shifts are multi-fold, with limited knowledge

about their types, often requiring expert insight to understand their nature. We present two real-data domain shift tasks for clinical event extraction, and analyze different types of realistic clinical domain shifts. Given that realistic domain shifts often have multifold reasons, we propose different methods based on domain shift assumptions, which can be further combined.

2.4 Longitudinal EHR Summarization and Finding Tracking

Methods of summarization and finding tracking can help understand longitudinal EHRs. Summarization methods can extract the most pertinent information from the lengthy clinical records to aid clinical decision-making. Finding tracking, similar to cross-document coreference resolution (CDCR) and clustering, can relate information across time that is worth comparing.

Summarization

LLMs have been applied to summarize the findings sections of individual radiology reports into impressions, effectively capturing critical findings. Van Veen et al. [97] apply LoRA fine-tuning to CLIN-T5-LARGE, and human evaluations show that the model generally captures critical findings but occasionally hallucinates details or infers prior history not in the report. Van Veen et al. [98] demonstrate GPT-4 with in-context examples can generate more complete summaries with fewer errors than expert-written impressions, according to human judgment. Zeng et al. [99] add in-context examples for Llama 3 70B, and observe improved automatic metrics such as ROUGE-L, but more factual errors according to human judgment. Revising these summaries with an LLM can correct presence errors but has minimal effect on errors related to progression status.

All the works mentioned above report automatic summarization metrics (BLEU, ROUGE-L, BERTScore) together with human evaluation, such as Likert-scale ratings on completeness and factuality [97], because those automatic summarization metrics can correlate poorly with human judgments [97]. For LLM text generation evaluation, fine-grained factuality evaluation systems [100, 101] are proposed to measure on atomic fact level. They decompose generated text into atomic facts and score each fact individually.

For longitudinal report summarization, Chien et al. [102] use LLMs to generate summaries directly from multiple reports. However, these unstructured summaries are difficult to verify, may omit findings and key details, and do not necessarily describe temporal changes for each finding. Structured summarization has been developed for heart failure [15], linking each fact in the summary to its source document and organizing facts into human-curated, disease-specific topics. However, fixed topic structures are not well suited for radiological findings, as groups of temporally related findings vary across patients.

Different from prior work, we develop a longitudinal report summarization approach to convert a patient's longitudinal chest imaging reports (CT, X-ray) into a timeline table, leveraging off-the-shelf LLMs and embedder models without using supervised labels. We first extract atomized lung-finding facts from the raw reports, then group temporally related facts to display the finding trajectories. Like some earlier work, we use ROUGE-L to evaluate findings, but we also provide scores for overgeneration of findings, quality of group names, and grouping of findings. Lastly, we compare automatic and human scoring for one LLM configuration.

Finding Matching

Finding matching tasks share similarity with CDCR and clustering tasks, and their related methods that inspire our approach.

CDCR entails detection of coreferring event mentions from multiple documents. Our lung radiological findings can be considered as events, but we group temporally related finding facts, instead of coreferring mentions. CDCR is often evaluated with same-document conference metrics, MUC, B^3, CEAF, their average CoNLL F1 [103]. Barhom et al. [104] develop supervised neural models that model mention pairs. Caciularu et al. [105] further improve the supervised results by leveraging cross-document contexts using Longformer-based models.

LLMs have been used directly for CDCR without supervised learning and require prompt engineering. Zhao et al. [106] use zero-shot GPT-4 to classify pairs of decontextualized sentences of event triggers that contain document-level information, outperforming untrained

crowd workers from MTurk. Min et al. [107] use GPT-4 to assign cluster indexes to mentions and output in a JSON format, and find that GPT-4 is worse than their supervised method, and using full contexts from all documents is worse than using only sentences containing the mentions. Sundar et al. [108] ask LLMs to replace placeholders next to the mentions with major entity tags in long narratives. When evaluated on gold mentions, GPT-4 is slightly better than their supervised method but not GPT3.5.

For radiology finding tracking, Datta et al. [109] present a CDCR dataset about tracking findings and devices across each patient's longitudinal radiology reports, and fine-tune a BERT model for pairwise mention coreference achieving low-to-moderate CoNLL F1 scores. Mathai et al. [110] use zero-shot LLMs to match a sentence in a follow up report to the single best matching sentence in its prior report.

The finding matching part in our structured longitudinal summarization is similar to the radiology CDCR task presented by Datta et al. [109], but the methods differs in that we do not group elements via pairwise associations. Instead, we create groups by mapping elements to a group label using LLMs, either via a question-answering prompt or using a prompt that inserts group tags for sequence of findings, inspired by Sundar et al. [108]. We also investigate the impact of full longitudinal contexts, given the mixed results in previous studies [105, 107].

Chapter 3

ROBUST SYMPTOM EVENT EXTRACTION USING ENCODER-ONLY LMS

In this chapter, we first introduce clinical event extraction and domain shift tasks in general. Then we describe cross-domain scenarios for extracting symptoms across institutions. Then we talk about the symptom event extraction under domain shifts. We propose domain generalization methods to improve robustness under domain shifts, including adaptive pretraining to improve clinical representations, and masking domain-specific symptom phrases to reduce overfitting to the training data.

3.1 Motivation for Symptom Extraction under Domain Shift

The assessment of symptoms, which are physical or mental problems that patients experience, is critical for disease diagnosis [111] and epidemic forecasting [112]. Symptoms are often documented in free-text clinical notes in Electronic Health Records (EHRs) and are not directly accessible as structured data for downstream applications such as disease prediction or healthcare outcome management. To address this need, a variety of entity and relation extraction methods [113, 1, 31, 3, 114] have been published to help automate the extraction of symptoms from clinical notes.

Clinical language varies across different institutions and specialties. Additionally, symptom phrase distribution shifts and symptom contexts vary across different patient populations. Considering the costs of manual annotations from medical experts, creating gold standard annotations to capture the entire clinical note distribution within EHRs does not scale. One common approach is to sample notes to annotate only from the domains of immediate interest (e.g., COVID-19 [3]) and apply the trained model to other target domains. In such cross-domain settings, we observed a significant extraction performance drop when a symptom extraction model trained on clinical notes of COVID patients is applied to clinical notes of cancer patients. [115]

Current state-of-the-art information extraction methods [1, 31, 32] use transformer-based language models (LM) [33, 116] that are pretrained on large scale corpora. However, the pretrained corpus may not cover the target task data distribution; for example, Bio+Clinical BERT [35] is trained based on clinical notes from patients admitted to intensive care units (ICUs) from one medical center. Our patient population is not limited to ICU patients and the notes are from two other medical centers. These types of data distribution differences are often referred to as "domain mismatch," but the term "domain" is not always well defined. For purposes of this work, a domain is defined as a collection of clinical notes associated with a particular set of patient cohorts, types of notes, and medical centers.

In order to avoid annotating training data in every new target domain, we train the model on the source domain where annotations exist and evaluate the trained model in the target domain. To improve the cross-domain performance, we improve the text representation by pretraining LM on relevant texts to our task based on the success of adaptive pretaining [88, 117]. Furthermore, we challenge the model by masking symptom phrases that are frequently seen in the source domain to reduce the influence of source domain features and force it to learn more generalizable contextual patterns.

In this work, we focus on cross-domain symptom event extraction from outpatient clinical notes from University of Washington Medical Center (UWMC) to cancer treatment ones from Seattle Cancer Care Alliance (SCCA). Our contributions are summarized below. First, we propose a domain generalization method that randomly masks frequent symptoms in the source domain during fine-tuning to encourage the model to give more attention to the context. Second, we observe that source-target domain differences impact the effects of both adaptive pretraining and symptom masking methods, and the source domain more distant to the target receives more benefits. Lastly, we find that masking frequent symptoms in the source domain helps detect symptom triggers[1] that are more likely labeled as non-triggers in the source domain.

[1]Triggers are symptom phrases themselves that indicate the events, and described by other arguments. More details in the Task section.

3.2 Task Definition

First, we introduce general event extraction tasks.

We extract structured clinical information as events from the sentences in clinical notes. Each event represents the occurrence of one piece of clinical information like a symptom or a radiology finding. Each event consists of a **trigger** that indicates the event occurrence and a set of **arguments** detailing its specifics. Both triggers and arguments are considered entities and are linked with relations. Each trigger or argument entity is associated with a **span**, and optionally with a **type** to classify their texts into categorical values.

The event extraction tasks are illustrated in Figure 3.1

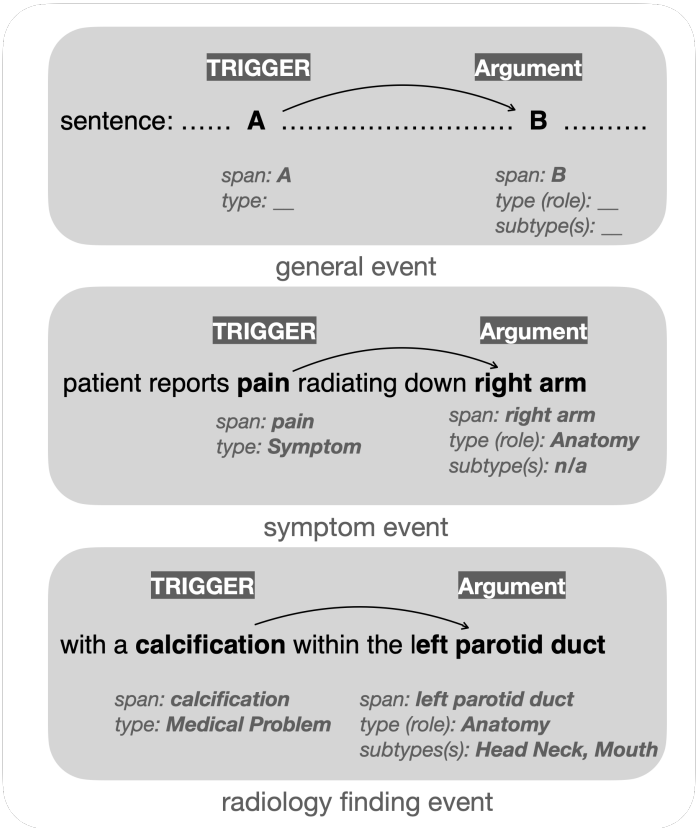

Figure 3.1: Clinical event schema. Upper: general event schema, Middle: symptom event schema, Lower: radiology finding event schema

During development, models undergo pretraining and SFT on labeled **source domain**.

Considering the costs associated with expert annotations on all domains of interest, as well as the efforts for updating pretraining, a model may need to process notes from a **target domain** distinct from the labeled source domain and possibly the domains of pretraining data.

In this thesis, we focus on **cross-domain** scenarios where adequate labeled data (at least a thousand sentences) from a source domain are available for SFT, and the evaluation is for the same IE task schema on a target domain different from the source.

In this chapter, we explore a symptom extraction task, where extracted symptoms are represented using an event-based annotation schema that is tailored to clinical text. We adopt the event schema from COVID-19 Annotated Clinical Text (CACT) Corpus [3], where **triggers** are symptom entities (e.g., pain, vomiting); argument entities are divided into **labeled arguments** (e.g., *Assertion*) and **span-only arguments** (e.g., *Anatomy*) depending on whether they contain entity subtype labels. Details of the symptom event schema are presented in Table 3.1 and example annotations are presented in Figure 3.2.

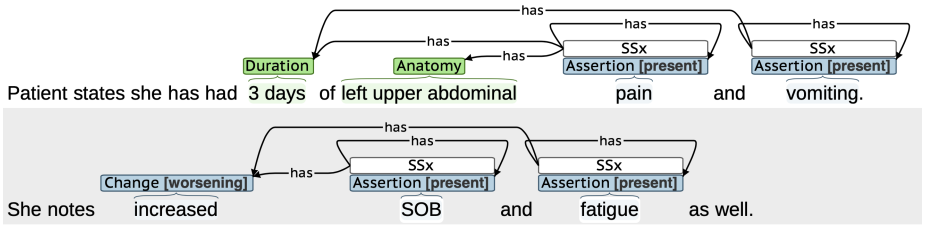

Figure 3.2: Annotated symptom event samples show entity spans with relations linking arguments to symptom triggers. Span colors denote symptoms (white), labeled arguments (blue), and span-only arguments (green). *Assertion* spans match triggers.

Our task is similar to joint entity and relation extraction tasks [118], but we require trigger entities to be present and arguments to be linked to triggers. We focus on span-level prediction; let $\mathbb{X}$ and $\mathbb{S}$ denote the space of all input sequences and their enumerated spans, and let Y_e and Y_r represent all entity types, and all relation types, including the negative types when labels are missing. For each input sentence x and its labels y_e and y_r, the entity classifier predicts the entity type of span s with $f_e : \mathbb{S}, \mathbb{X} \rightarrow \mathbb{Y}_e$, where $f_e(s, x) = y_e(s)$, and

Event type,	**Argument type**	**Argument subtypes**	**Span examples**
Symptom	SSx - Trigger*	–	“cough," “shortness of breath"
	Assertion*	{present, absent, possible, conditional, hypothetical, not patient}	“admits," “denies"
	Change	{no change, worsened, improved, resolved}	“improved," “continues"
	Severity	{mild, moderate, severe}	“mild," “required ventilation"
	Anatomy	–	“chest wall," “lower back"
	Characteristics	–	“wet productive," “diffuse"
	Duration	–	“for two days," “1 week"
	Frequency	–	“occasional," “chronic"

Table 3.1: Entity types and subtypes for symptom events. Entities can be grouped into triggers, labeled arguments, and span-only arguments. Labeled arguments, including *Assertion*, *Change* and *Severity*, have subtypes [3].

the relation classifier predicts whether span s and span s' are linked using $f_r : (S, S) \rightarrow \mathbb{Y}_r$ where $f_r(s, s', x) = y_r(s, s')$.

After prediction, events are constructed for each trigger entity, adding non-trigger entities as event arguments if they are linked to the trigger head.

The *Assertion* argument includes subtype labels that indicate whether the identified symptom is *present*, *absent*, *conditional*, etc. For *absent* symptoms, there are consistent negation cues, like "denies" or "no." While there are affirming cues, like "reports" or "has" for *present* symptoms, the *present* subtype is often implied by a lack of negation cues. To provide the *Assertion* span classifier with a more consistent span representation, we replaced each *Assertion* span with the trigger span in each event. The extraction model treats each trigger and its *Assertion* argument as the same entity which has the trigger span and the *Assertion* subtype. We unmerged triggers and *Assertion* arguments into entities with the same spans in post-processing for evaluation purpose; *Assertion* entities keep the subtype labels, and trigger entities do not.

3.3 Datasets with Domain Shifts

In this work, we use the following four resources.

The COVID-19 Annotated Text Corpus (*COVID*) contains clinical notes of 230,000 outpatient clinic patients treated at UWMC between May and June 2020 [3]. *COVID* contains telephone encounter notes, progress notes, and emergency department notes from UWMC.

The Lung Cancer Annotated Text Corpus (*Lung*) [115] was constructed from clinical notes of 4,673 lung cancer patients diagnosed between 2012 and 2020. *Lung* includes outpatient progress notes, admission notes, emergency department notes, and discharge notes that were created 24 months prior to cancer diagnosis at UWMC.

The Ovarian Cancer Annotated Text Corpus (*Ovarian*) [115] is based on the clinical notes of 173 ovarian cancer patients diagnosed between 2012-2021. *Ovarian* includes outpatient progress notes, admission notes, emergency department notes, discharge notes, and gynecology notes created 12 months prior to cancer diagnosis at UWMC.

The Post-Cancer Diagnosis Annotated Text (*Post*) was built using clinical notes from

2,968 prostate cancer patients and 1,222 Diffuse Large B cell lymphoma patients diagnosed between 2007-2021. *Post* contains notes from SCCA after cancer diagnosis, including notes from urology, oncology, hematology, surgery, radiation oncology, and palliative care.

All four resources include a variety of clinical notes from UWMC and SCCA. Lung, Ovarian and Post datasets are all related to cancer patients. Lung and Ovarian share the same group of annotators. COVID has the most labeled examples. For each domain, we retrieved unlabeled clinical notes from the same distribution as the labeled ones. Table 3.2 provides more details about the data size, with inter annotator agreement for the three most frequent entity types.

Source	Labeled Reports	Labeled Sentences	Trigger Instances	Inter-Annotator Tr/As/An	Unabeled Reports
COVID	1028	89573	16966	86/83/81	87723
Lung	145	22190	3937	74-83/70-79/71-79	281493
Ovarian	100	15031	2477	79-82/71-81/64-79	29298
Post (Test)	200	28764	5619	80-83/73-74/74-76	–

Table 3.2: Dataset information for labeled training data in the source domains, labeled test data in the target domain, and unlabeled data in both the source and target domains. Interannotator agreement range is given for Triggers (Tr), Assertions (As), and Anatomy (An).

In this study, we choose Post as the target domain, and we use two source sets: COVID and the combination of Lung & Ovarian (due to their small size). All source domains differ from the target in institution as well as diagnosis stage and diseases types of patients. The Lung & Ovarian source is similar to the target domain as it also involves cancer patients, but it does not relate to prostate or lymphoma cancers. The impact of these domain differences between source and target domains is illustrated in Figure 3.3, which plots the coverage of triggers across domains.

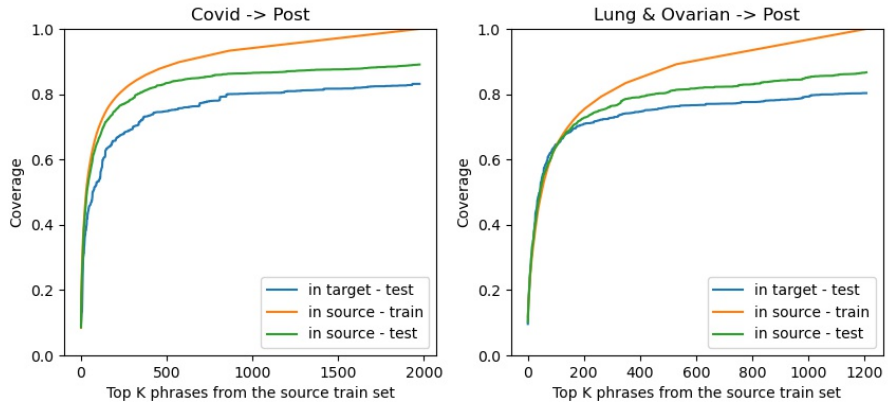

Figure 3.3: Domain discrepancy: source-target domain differences in terms of trigger coverage.

3.4 Generalization Methods for Encoder-Only Models

In our cross-domain setting, we assume labeled training data is available for the source domain only, and that the source differs from the target domain. The baseline cross-domain model leverages a standard configuration of a pretrained transformer incorporated into a model that is trained on the labeled source domain data. The experiments look at two strategies for improving cross-domain performance: (i) the use of unlabeled data in the source domain for adaptive pretraining, and (ii) the use of dynamic masking in the supervised training stage. We explore the benefits of the two approaches alone and in combination for different degrees of domain mismatch. Both adaptive pretraining and supervised training are on source domain data; inference at test time is on target domain data.

3.4.1 Baseline

All experiments are run on the transformer-based SpERT [1], which is one of state-of-art models on the SciERC benchmark [118]. According to the original SpERT work [1] and a clinical application based on SpERT [2], the model architecture is shown in Figure 3.4. SpERT contains light-weighted entity and relation classifiers on top of the BERT encoder [33]. Specifically, we use a clinical version of BERT, Bio+Clinical BERT [35]. The **entity**

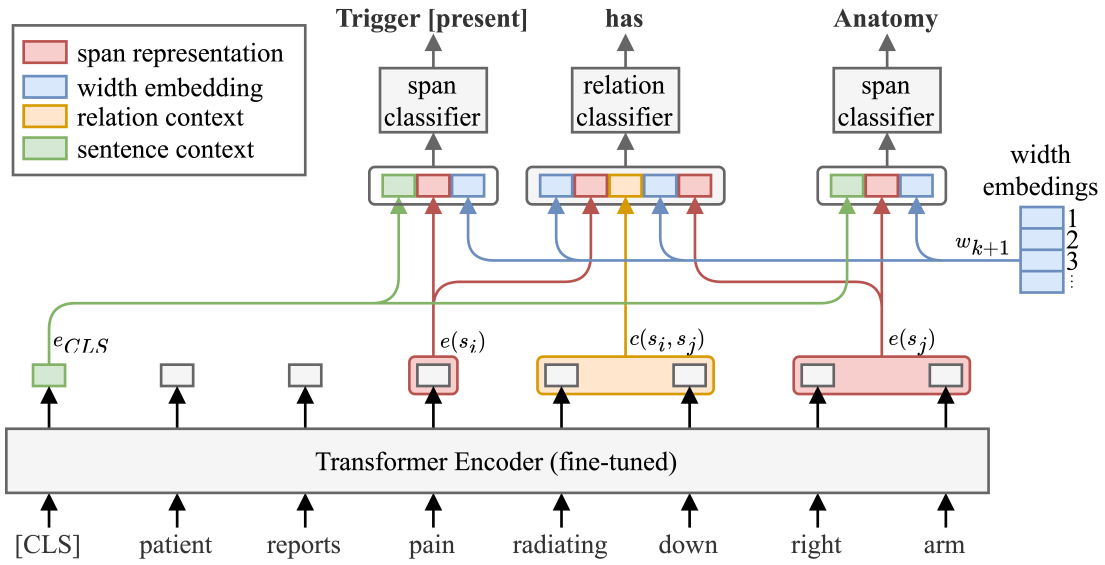

Figure 3.4: SpERT Model Achitecture based on work [1, 2]

classifier predicts enumerated spans at the subtype level. To represent each span, SpERT groups the encoder outputs of tokens within the span by max pooling, and concatenates it with the sentence context and span size embeddings. To predict whether a relation exists between any pair of entity spans, we represent the span pair using their entity representations plus the max pooled encoder vectors of the words between the two spans. The **relation classifier** classifies the relation representations into a binary outcomes.

To train the model, we minimize the cross entropy (CE) loss between predictions and labels, sampling non-entity spans and span-pairs without relations as negative examples. To normalize the outputs of the multiclass entity classifier, we use the softmax function, while for the binary relation classifier, we use sigmoid. For input sentence x, labels for entities y_e and relations y_r, let f_e and f_r denotes the model outputs of entities and relations, the objective is $L_{Joint} = L_{Entity} + L_{Relation}$, where

$$L_{\text{Entity}} = \sum_{s \in Spans(x)} CE(f_e(s, x), y_e(s))$$

$$L_{\text{Relation}} = \sum_{(s,s') \in SpanPairs(x)} CE(f_r(s, s', x), y_r(s, s'))$$

3.4.2 Adaptive Pretraining

The baseline model is initialized with Bio+Clinical BERT [35] pretrained on biomedical research articles and clinical notes. However, the pretraining corpus could fail to represent our task distribution, because its clinical note source, MIMIC-III [87] is mostly from ICU patients from one medical center, but we are targeting cancer patients who are not necessarily admitted to ICUs and from another medical center. To adjust the LM to the target domain, we continue pretraining on unlabeled texts relevant to our task as in the DAPT method [88]. Pretraining minimizes the masked language model (MLM) objective. For any input x with tokens $t_0, t_1, ...t_l$, we randomly replace tokens t_i to [MASK] at a fixed rate (15%). Let the set of masked token indexes in sentence x be $LMMask(x)$ and the sentence after masking be $\tilde{x}$; let the LM $f_{LM}(i, \tilde{x})$ predicts for the original i-th token that is masked in the new context $\tilde{x}$, the MLM pretraining objective is

$$L_{\text{MLM}} = \sum_{i \in LMMask(x)} CE(f_{LM}(i, \tilde{x}), t_i)$$

In adaptive pretraining, we use the combined set of unlabeled texts from COVID, Lung, and Ovarian, with domain indicator tokens at the start of a sequence. We also conducted experiments using unlabeled target domain texts as an additional pretraining corpus, but results are not reported since it was not helpful. We adopt the same masking rate of 15% for the MLM pre-training objective as in the original BERT model [33].

3.4.3 Dynamic Masking

As shown earlier, domains may differ in the frequency of specific trigger words. We hypothesize that, for very frequent triggers, the model may put too much weight on the word sequence alone vs. its context, failing to capture more generalizable contextual cues. In this scenario, cross-domain performance might be compromised, especially when symptom phrase distribution shifts and symptom contexts change. The model may fail to detect symptoms with a considerable change in context distributions, as well as symptoms less frequent in the source.

We challenge the model to learn the contextual information for trigger extraction by masking the frequent symptom phrases seen in training. This is a data manipulation method

that changes the labeled source data without modifying the model. First, we create the symptom phrase list based on their frequency ranking in the source domain train set, then search those phrases in the source domain training data, and randomly mask the matched phrases with a fixed probability, regardless of whether they are annotated as triggers or not. We treated the random masking rate as a tunable hyperparameter and determined the optimal rate for our task is 80%. In order to prevent permanent loss of annotations due to aggressive masking, we use dynamic masking [116] to change random masks every epoch. We train the joint entity and relation model on the masked source domain labeled data using the same L_{joint} objective.

3.5 Experiments and Results

3.5.1 Implementation

Each note was split into sentences using spaCy.[2] To encode the text, we use Bio + Clinical BERT's default tokenizer, which contains 28996 uncased tokens. As in Bio + Clinical BERT, the transformer encoder contains 12 layers and 12 heads. The training batch size is 15. For each model, we train it for 10 epochs.

During pretraining, we truncate the text corpus into chunks of 512 length. For each optimization step, we accumulate sub-batches of size 32 eight times in order to have a larger batch size of 256. The MLM random rate is 15%. When pretraining on all source domains together, the LM is trained for 31 epochs finishing 67.6 K steps. With the target domain texts added, we pretrain on this larger corpus for 30 epochs, which is equivalent to 126.8 K steps. To create the symptom phrase lists for masking, we used the top 200 frequent symptom phrases from each source. In order to simplify the list, we keep only phrases with one token and exclude punctuation-only[3] or single-character triggers. In training data, we search for those listed tokens, and dynamically mask matched tokens at an 80% random rate.

All SpERT modeling experiments are conducted on a NVIDA GoForce GPU card with

[2]https://spacy.io/models/en#en_core_web_sm

[3]Trigger labels on punctuation could come from annotation errors.

11.6 GB of memory. A NVIDA A100 GPU with 42.5 GB of memory is used to pretrain the language models.

3.5.2 Evaluation

The evaluation process is similar to our previous symptom event extraction work [3]. All triggers and arguments are scored by micro F1 scores. The predicted trigger is considered correct if the trigger type and span are both equal to the gold trigger. The subtype-level entity label and the linked trigger must match the gold labels for a predicted labeled argument to be considered correct. When calculating span-only argument metrics, we count the overlap tokens between each predicted span and the gold span, and we require the two spans to have the same entity type and trigger. When comparing models, we run 5 random seeds and report the results of a two-sided T-test.

3.6 Results & Discussion

Table 3.3 summarizes the cross-domain results on the trigger and argument extraction. The Lung & Ovarian source domain performs better than the COVID domain for the three most frequent entity types.

Adaptive pretraining benefits performance for both source domains. For the COVID source model, the gains are significant for trigger (*SSx*), *Assertion* and *Characteristics*. For the models trained on the Lung & Ovarian source, The gains are significant in *Anatomy* and *Frequency* arguments.

Dynamic masking benefits performance for the models trained on the COVID domain, but has varying effects for different source domains. Specifically, dynamic masking improves the trigger extraction of COVID models with statistical significance, for both scenarios with and without adaptive pretraining, but hurts the Lung & Ovarian source model performance when no adaptive pretraining is applied. The improvement for COVID models on trigger F1 scores is primarily driven by the increase in recall. There is a trade-off between recall increase and precision decrease, but recall dominates the change. However, for Lung & Ovarian source models, without adaptive pretraining, the precision of trigger extraction significantly decreases after dynamic masking. The varying effects can likely

Entity	NT	COVID				Lung & Ovarian			
		Baseline	w/ Mask	+ Adapt Pretrain	w/ Mask	Baseline	w/ Mask	+ Adapt Pretrain	w/ Mask
SSx	5629	77.2	78.5**	78.6*	**79.9***	79.2	78.2- -	**79.3**	79.1
Assertion	5629	72.2	73.1*	74.0**	**74.9**	**75.1**	74.4-	75.0	**75.1**
Change	401	55.7	**56.9**	**56.9**	**56.9**	47.2	44.8	46.1	**47.3**
Severity	291	30.7	30.4	29.6	**33.3***	40.7	36.4- -	**42.2**	38.5
Anatomy	3007	55.3	56.7	57.2	**59.4**	60.7	58.5- -	**63.2***	61.4
Characteristics	1250	22.5	22.8	**26.8****	25.8	15.1	16.6	16.3	**17.7**
Duration	837	46.5	44.1	46.3	**48.4**	29.9	32.5	30.4	**34.2**
Frequency	301	45.9	45.6	44.9	**46.6**	24.7	24.9	**27.3***	26.6
SSx (Precision)		**90.1**	89.4- -	89.4- -	88.9	**84.1**	82.5-	83.1	83.0
SSx (Recall)		67.5	69.9**	70.1*	**72.6***	74.8	74.3	**75.9***	75.6

Table 3.3: Entity micro F1 scores and trigger (SSx) precision and recall scores for both sources. **(- -) and *(-) indicate significant gain(loss) relative to the approach before adding the lastly applied domain generalization strategy at p-value smaller than 0.01 and 0.05, respectively. We have four different versions of models: Baseline, Baseline with dynamic masking, Adaptive Pretrainining, and Adaptive Pretraining with dynamic masking. "NT" stands for "number of true labels," which refers to the count of gold labels.

be attributed to differences in the distributions of symptom phrases and symptom trigger contexts across domains.

The combination of masking and pretraining makes it possible to take advantage of the greater amount of labeled data, giving performance that is similar to or better than the best result from the Lung & Ovarian source model for all entity types except *Severity* and *Anatomy*. The best COVID source model gives a result close to the human annotator agreement for triggers (80-83) and *Assertion* (73-74) entities.

In order to better understand the effect of trigger frequency in the source domain, we looked at performance for different sets of triggers grouped by source frequency. Table 3.4 provides F1 results for the different training configurations. For COVID, masking has little impact on the performance for the most frequent source triggers, but significantly benefits the triggers in the top 20-40 and top 60-80 groups. In contrast, masking significantly hurts

Source	Version	Top 20	Top 20-40	Top 40-60	Top 60-80	Top 80-100
COVID	Baseline	95.4	89.9	83.9	83.1	82.6
	w/ masking	95.7	**93.1****	84.2	87.1**	82.7
	+ Adapt Pretrain	96.3	90.2	84.2	82.6	**83.8**
	w/ masking	**96.5**	92.8**	**85.8***	**87.2***	83.0
Lung & Ovarian	Baseline	**97.4**	77.8	81.9	**76.6**	73.8
	w/ masking	96.2- -	77.5	**82.5**	72.9- -	73.6
	+ Adapt Pretrain	**97.4**	77.4	80.6	76.2	74.8
	w/ masking	96.6- -	**78.3**	82.3	73.6 -	**76.0**

Table 3.4: Trigger F1 results on subsets of symptom phrases grouped by source domain frequency. **(- -) and *(-) indicate significant gain(loss) relative to contrasting condition at p-value smaller than 0.01 and 0.05, respectively.

the most frequent source triggers when the source is Lung & Ovarian.

Masking benefits phrases that are are more likely associated with negative contexts in the source than in the target. A trigger phrase that is frequently annotated as a non-trigger in the source can lead to false negatives in the target domain. To study how the negative context shift affects masking, we plot positive class ratio of trigger phrases on both source and target domains, which is the likelihood of the phrases annotated as triggers (Figure 3.5). We size and color each point according to the absolute value and sign of the false negative change. We show only false negatives because the trigger detection improvement after masking is due to the reduced false negatives. Most phrases with reduced false negatives are above the diagonal, in other words, the phrases improved after masking are more likely to be annotated as non-triggers in the source domain than in the target. The COVID source has more trigger phrases overly associated with negative contexts and sitting above the diagonal, which can be compensated for by masking.

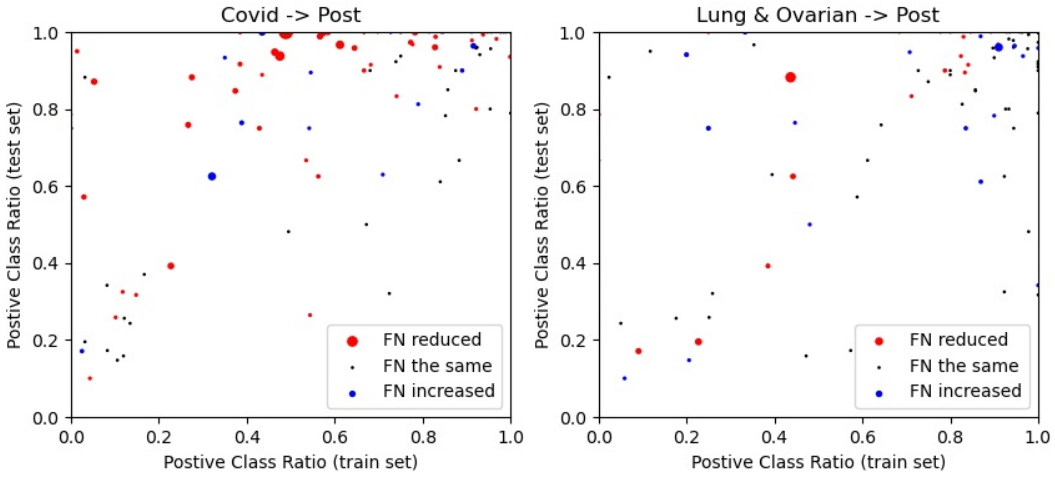

Figure 3.5: Relative frequency of a positive trigger label in the test vs. training set for the 100 most frequent triggers in the test set. Size of the points indicates the absolute value of the symptom phrase false negative (FN) change after masking. Red indicates that false negatives are reduced; blue indicates no change or an increase.

3.7 Conclusion

We use adaptive pretraining and our proposed dynamic masking method to improve symptom event extraction under cross-domain settings without using additional labeled data. Both methods significantly benefit when the source domain is more distant from the target, and achieve trigger F1 scores close to the human annotator agreement. The dynamic masking method improves the detection of symptoms that are less likely to be annotated as symptom triggers in the source domain. For future work, understanding the type of domain discrepancy could guide us in selecting domain generalization (or adaptation) methods.

While we primarily focus on analyzing the symptom phrase characteristics when describing the shifts, other factors such as the amount of source data, and annotation consistency can also affect domain generalization. Future work should explore ways to analyze and handle different types of shifts, and we recommend cross-domain studies controlling other factors such as annotation quality.

The limitation of this work is that the entity-type labels are encoded as one-hot encodings, thus each label's representation depends on the data points from this category, putting

minority categories in disadvantage. One extension is to represent the labels in texts that describe the labels in natural language. In the next chapter, we will introduce text-to-text QA event extraction as a more generalizable event extraction under label imbalance.

Chapter 4

ROBUST RADIOLOGICAL FINDING EXTRACTION WITH GENERATIVE LMS

This chapter focuses on domain generalization for radiology finding extraction tasks when using generative LMs. I first describe the radiological finding extraction task, and the domain shift scenarios across exam modalities. The domain generalization methods include breaking down the complex tasks in finetuning to reduce the compositionality restricted to the source domain, and retrieving target-domain contexts to enhance clinical knowledge. Concerning computation costs in clinical applications, we only use moderately-sized LMs.

4.1 Motivation

Radiology reports contain a diverse and rich set of clinical abnormalities documented by radiologists during their interpretation of the images. Automatic extraction of radiological findings would enable a wide range of secondary use applications to support diagnosis, triage, outcomes prediction, and clinical research [119]. We adopt an event-based schema to capture both indications, the reason for radiology exams, and abnormal findings documented in radiology reports. We use an annotated a corpus of reports from three distinct radiology examination modalities [9]: Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), and Computed Tomography (CT). Each event consists of a trigger, words that indicate a particular indication or finding (e.g., lesion), and a set of attributes (assertion, anatomy, characteristics, size, size trend, size count) that describe this indication or finding. Manual annotation of radiology reports is costly, therefore we hope models can generalize across different exam modalities. In this work, we define each modality in our annotated corpus as a domain and study cross-domain generalization among different modalities for the task of event extraction. Event extraction can be conceptualized as a series of subtasks, which include entity detection (trigger and attribute spans), relation detection (between triggers and attributes), and entity normalization (fine-grained labels on

spans). In our experiments, we focus on trigger detection and anatomy attribute extraction with normalized labels.

To enhance generalization capabilities, some studies employ generative models and formulate tasks as question answering and using texts to represent both inputs and outputs [39, 41], as opposed to allowing the model to solely learn task intent from training data [1, 4].

The exceptional performance of generative models often rely on large model size; however, in real-time inference for processing large-scale clinical notes, reducing inference costs is crucial. To address this need, for task inference, we want to reduce the number of decoding passes and employ smaller models. Due to the high inference costs, there is a desire to merge these subtasks and decode them in a single step. However, the generative approach has been reported to perform better on solving subtasks individually but worsen when combined, a phenomenon referred to as the compositionality gap [82]. This gap can be exacerbated under domain shifts when models learn subtasks jointly, as interdependence of subtasks may vary across domains.

While large language models (LLMs) mitigate the compositionality gap using reasoning steps [83, 82] to solve complex questions by decomposing them into smaller ones, there is limited work on reasoning for highly specialized domains (such as medical event extraction) or with smaller models. In this paper, we reduce the compositionality gap for smaller models through formatting of complex tasks into easier subtasks as blocks. This approach teaches models how to solve individual subtasks independently and how to assemble them for solving more complex tasks.

The generative model enables seamless integration of supplementary contexts into the prompt, which compensates for the knowledge gap to larger models and reduces inference costs. To aid in domain adaptation, we extract target domain contexts that are likely to be helpful for the task, instead of retrieving similar contexts for general purpose. Specifically, to assist with anatomy normalization tasks, we employ an unsupervised extractor to acquire pertinent contexts that likely contain anatomical information from the same document and/or unannotated text from the same domain. This process can either disambiguate the original single-sentence input or provide anatomy-related hints that the model can utilize.

To avoid introducing source-domain-specific reliance on the contexts, we incorporate the contexts only at the inference stage.

In our experiments, we first study domain shift for extracting radiology finding events and observe that cross-domain performance decline is more pronounced for knowledge-intensive anatomy normalization tasks, while detecting entity spans exhibits relatively stable performance. We demonstrate that building subtask blocks and assembling them as sequences to solve complex tasks can reduce the compositionality gap in smaller models. We show that incorporating target-domain contexts in domain adaptation can compensate for reduced model sizes, enabling good performance with smaller models.

4.2 Task

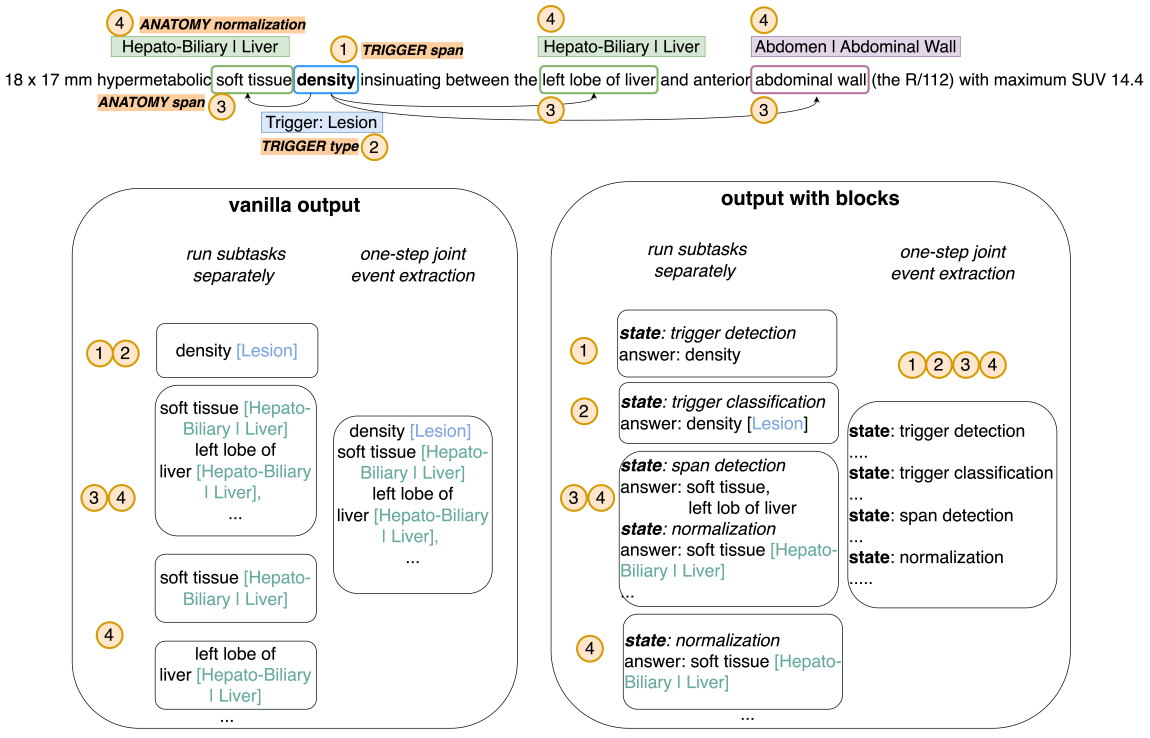

Figure 4.1: Representations of anatomical information in radiology reports, with the event-based annotation at the top and two generative model output formats to multi-step and one-step processing. The left-hand side shows the **vanilla format** and the right-hand side shows the **building block format**.

4.2.1 Event Extraction for Radiology Findings

Our event schema includes three event types: i) *Indication* is the reason for the imaging (e.g. motor vehicle accident or cancer staging); ii) *Lesion* captures lesions uncovered by the exam (e.g. mass or tumor); and iii) *Medical Problem* characterizes non-lesion abnormalities (e.g. fracture or hernia). Each finding event is characterized by an event trigger and set of attributes (assertion, anatomy, characteristics, size, size-trend, count). In this work, we focus only on extracting events with normalized anatomical information and investigate cross-domain generalization for different exam modalities. Figure 4.1 presents a *Lesion* event example. The event extraction process can be broken down into four subtasks: (1) Trigger span extraction (e.g., "density"), (2) Trigger type classification (e.g., "density" - Lesion), (3) Anatomy span extraction (e.g., "left lobe of liver" associated with the trigger "density"), and (4) Anatomy normalization to parent-child anatomy categories (e.g., "left lobe of liver" - Parent: Hepato-Biliary, Child - Liver). See Appendix A.1 for the full list of hierarchical parent-child anatomy categories.

We evaluate event extraction performance using the F1 metrics by Lybarger et al. [2]. Our assessment of the trigger extraction is based on the span overlap and the event type match with respect to the gold standard labels. The anatomy extraction is first assessed at the span level. A correct anatomy prediction is associated with a correct predicted trigger and anatomy span overlap with the gold standard labels. Additionally, we evaluate anatomy extraction based on the normalization level, irrespective of their spans. A match between the predicted anatomy entity and the gold label indicates that the trigger is matched, and the normalized anatomy category is equal.

4.2.2 Domain Shifts across Radiology Modalities

Our research investigates cross-domain generalization among three distinct radiology exam modalities: MRI, PET, and CT. These exam modalities are performed for different reasons with different technologies and the resulting radiology reports differ in terms of level of details as well as anatomy distribution. While CT and MRI scans allow radiologists to view structures inside the body, a PET scan, on the other hand, captures how tissues in the body

work on the cellular level and shows unusual activity. MRI scans very frequently involve neurological exams. The most common use of PET scans is to diagnose or monitor certain cancer types. In our experiments, we define each modality as a domain. We use PET as the target domain, and train on three domains separately to evaluate both in-domain and cross-domain scenarios.

4.3 Generalization Methods for Encoder-decoder Models

4.3.1 Generative Event Extraction with T5

In order to improve the model's generalization capabilities over BERT-based alternatives [4, 1], we structure our event extraction task in a unified question-answering (QA) format [41, 39]. With the generative approach, the model leverages the semantic meaning of prompts for specifying subtasks and associated categorization labels. Based on experiments with in-context learning [43], we expect this to be beneficial for domain-mismatches in class label distributions, e.g. where infrequent classes in the source domain are frequent in the target domain. Furthermore, the text-to-text format offers the flexibility to incorporate additional contexts to facilitate tasks, as discussed in Section 4.3.3.

The input prompt comprises: (1) an input sentence from clinical notes to extract events from, (2) a question that describes the task or subtask, and (3) an ontology that provides textual labels for classification tasks and hierarchical relationships if multi-level granularities are required. The output is a word sequence that specifies the extracted information (the answer). Two alternative output formats are discussed in the next section; example input-output pairs for both are in Appendix A.2.

Event extraction can be seen as a multi-hop question-answering process, involving a series of subtasks for successful completion. We use a pipeline approach to address the event extraction subtasks in different steps, where each step in the pipeline consists of a specialized generative model trained for one or more of the subtask types. Three different architectures are explored:

Three-step approach: This involves a first step for detecting trigger spans and trigger types, followed by a second step for identifying the anatomy associated with each detected

trigger, and a third step for normalizing each identified anatomical entity at parent and child levels individually.

Two-step approach: This involves a first step for detecting trigger spans and trigger types, followed by a second step for identifying and normalizing the anatomy associated with each detected trigger.[1]

One-step approach: we address all subtasks, which may be associated with multiple entities, in a single pass per input sentence. This method results in longer output lengths compared to the individual steps of previous two approaches.

The one-step approach substantially reduces inference costs compared to other two multi-step approaches. However, we find that it negatively impacts model performance due to the longer output and the compositionality gap. The performance loss is mostly recovered by changing the output format (as described next) together with a multi-task training strategy. Specifically, we train the model on both the complete task and the decomposed subtasks. This allows the model to perform subtasks independently and assemble subtask sequences for complex tasks. During inference, we decode in a single step to minimize costs.

Our work builds on generative models, specifically the clinical version of the T5 models [5], which are pre-trained on medical articles and clinical notes. This choice leverages their strengths in comprehending clinical text styles and medical knowledge.

4.3.2 Output Formats

We explore two different output formats as illustrated in Figure 4.1, with subtask answers provided in sequence when there are multiple subtasks.

The baselines leverage a standard output format (referred to here as the **vanilla format**), which specifies the answer for a subtask with an extracted span followed by the entity label in brackets "[]". When multiple entities are detected, they are generated in sequence.

The vanilla format can be used with the one-step approach, but the resulting output can be very long when multiple triggers and/or entities are detected. The lack of distinction between types of spans in the output makes it harder for the language model to learn

[1]Both the 2-step and 3-step approaches use the same second step, predicting anatomy spans and their normalized values. The three-step approach drops the normalized values from its second step.

the subtask structure. To address this problem, we introduce a state-augmented prompt (referred to as the **building block** format), in which each subtask is associated with a state (as in a finite-state transducer) and explicitly named. Our approach is motivated by the work on chain-of-thought LLMs [83, 82], which use natural language reasoning in the generated outputs to address the compositionality gap. However, it differs in that we do not use natural language reasoning, but rather more of a programming-like description. In addition, the finite-state framework is amenable to meulti-task training, which is particularly important for the block approach.

4.3.3 Using Target-domain Contexts in Prompts

A single input sentence may not provide enough information for a model to complete a task, as additional details may be needed for disambiguation or to supplement missing knowledge in pre-trained language models. Fortunately, the text format of the input allows for the seamless integration of additional contexts from the target domain during inference to aid in the task and infuse helpful domain-specific bias, even if the models were not trained for reading contexts.

The desired contexts should be relevant to the input sentence and contain helpful task information. We utilize two types of contexts: document-level and domain-level contexts to help anatomy normalization subtasks. Document-level contexts include adjacent sentences before and after the input, automatically extracted section headers[2] and exam type metadata associated with the same clinical note. The document-level contexts are likely to describe relevant anatomical parts, as section headers and exam types often summarize anatomical information. Domain-level contexts are retrieved from the unlabeled target-domain corpus. We search for the most similar sentence with the greatest lexical overlap degree, using the search algorithm BM25 [120].[3] When the search pool is large, the top-ranked retrieved context sentence likely describes a similar anatomy part as the queried input sentence. To reduce computational costs and ensure that the retrieved sentences contain useful anatomical

[2]We extract section headers as the beginning of the last previous sentence containing ':'

[3]We implement the BM25 algorithm using https://github.com/dorianbrown/rank$_b m25$

information, we pre-filter the target corpus to limit the search scope to sentences containing common anatomy terms listed from anatomy normalization categories and high-frequency auto-extracted section headers, reducing the number by 74%. More context-retrieval details are in Appendix A.4.

We add contexts only during decoding (and not in training) to prevent the model from relying too much on source-domain contexts. In the input prompts, exam type, section headers and prior sentences are placed before input sentences, following their natural orders. Other contexts are inserted between the input sentences and task ontology.[4] We test this approach in a separate anatomy normalization run after the one-step building block model. This process combines building block output format with target domain context integration. The reason for not directly adding it to a one-step process is that introducing contexts to inputs can potentially corrupt span detection, as the model may extract spans from the context rather than exclusively from the input sentence.

4.4 Experiments

4.4.1 Radiology datasets across exam modalities

Data split	Note Count	Sent Count
CT (train)	143	3707
MRI (train)	144	3551
PET (train)	142	5184
PET (valid)	20	758
PET (test)	40	1481
PET (unlabeled)	1471	50000

Table 4.1: Dataset statistics for the three radiology examination modalities: CT, MRI, and PET. We explore in-domain and cross-domain training, evaluating on PET.

We use an annotated corpus containing radiology notes about CT, MRI, and PET

[4]The full T5 input template is described in Table A.2 from Appendix A.2

imaging exams; statistics are given in Table 4.1. The anatomy normalization labels are grouped into sublevels according to the SNOMED CT concepts [121]. Notes in the test and validation sets are all doubly annotated. The inter-rater agreement for Trigger is 0.73 F1.

Variations in anatomy distribution across imaging modalities can cause domain discrepancies. PET has the most balanced distribution among parent-level anatomy categories, followed by CT. However, MRI has a heavily skewed distribution, with 62% of trigger-associated anatomy entities being neurological among 16 parent-level categories. See Appendix A.1 for anatomy distribution details.

To enhance domain-specific context retrieval and boost the chances of retrieving helpful contexts, we expand the search pool by sampling 50,000 unlabeled PET report sentences from the same distribution as in the annotated reports [9], with a minimum of three tokens.

4.4.2 Implementation

In the non-generative baseline, we adopt the mSpERT model [4] for hierarchical multi-label entity and relation extraction. Entities are extracted as spans. We initialize with Bio-Clinical BERT [35].

For the T5 model using both vanilla output formats and the subtask block formats, we initialize with ClinicalT5 [5]. Details of the post-processing applied to T5 outputs for obtaining entity spans and types are provided in Appendix A.3.

For all models, the best checkpoint is chosen after 15 training epochs based on the validation performance on the target domain. For T5 models with multitask training on subtask blocks, which involves a higher number of training steps, we evaluate the model on the validation set after every 0.5 epoch approximately. For methods that do not involve multitask training, we evaluate the model on the validation set per epoch.

We implement multitask training on subtask blocks for MRI and PET, using the auxiliary tasks, as described in Section 4.3.1, including trigger span detection, trigger classification, joint anatomy span detection and normalization, and anatomy normalization. For the CT-PET transfer scenario, we add an additional anatomy span detection auxiliary task, as we observe that more aggressive learning is needed for anatomy span detection in the CT

domain. Detailed information about hyperparameters can be found in Appendix A.5.

4.5 Results

Table 4.2 shows the trigger and anatomy detection results for mSpERT compared to different context-independent T5-base alternatives. For the in-domain condition, all T5 approaches outperform the mSpERT model for the three anatomy-related metrics. The results for trigger detection are mixed, but fairly similar for all. The best performance overall is obtained using the 2-step vanilla output T5 model. For the cross-domain scenarios, all models suffer degradation in performance compared to the in-domain condition, with the greatest performance drop for the normalized anatomy categories, particularly for the MRI-PET condition which has the greatest mismatch in anatomy distribution. The performance loss is greatest for the mSpERT model, with a 44% relative reduction in F1 scores for normalized anatomy (at both parent and child levels) for the MRI-PET case. In contrast, the relative loss on the parent and child levels for the T5 models is 24-29%. For both within and across-domain scenarios, the building block technique improves the 1-step results for all categories, but particularly for the more difficult anatomy normalization tasks. As described later in Section 4.6.2, the 1-step approach is sensitive to the compositionality gap, which is ameliorated by the block approach. For the cross-domain scenarios, the best overall results are obtained with the 3-step approach for the CT-PET condition and with the 1-step block approach for the MRI-PET condition (greater mismatch). An additional advantage of the 1-step approach is the lower latency associated with using only one decoding pass.

As described earlier, target-domain contexts are added to prompts during a second step of T5 decoding to help anatomy normalization, after the 1-step subtask block decoding with T5-base. Table 4.3 shows results for all different types of contexts, as well as using either T5-large or T5-base in the second step without context. Without context, the T5-base and T5-large models give similar results for in-domain and CT-PET cross-domain conditions, but T5-large improves results for the MRI-PET condition. (Note that T5-large is only used in the last step; a bigger benefit could be observed if used in both steps.) All types of context are useful for the two domain-shift cases, but there is little or no benefit for the in-domain case. Of the different types of context, automatically retrieved similar sentences

from unlabeled target-domain data provide the greatest benefit in the mismatched scenarios. Combining all contexts provides a small additional benefit, except for the anatomy parent in the MRI-PET case. Anecdotally, we observe that same-document contexts are useful for disambiguation, while hints for challenging examples are more likely collected from a large domain-level corpus rather than just the same document. (For examples, see Appendix A.6.)

Table 4.4 provides information on the relative cost of the different T5 models. The multi-pass models have higher latency (average passes/sample) in that passes are necessarily sequential. (Note that samples with no findings or no anatomy identified in the first pass do not require additional passes.) The number of tokens per sample is an indicator of cost. The 1-step model with blocks has a higher cost than the 2-step approach because of the additional tokens introduced by the state-augmented prompt, but the cost is still lower than the 3-step approach. The use of context adds additional cost.

4.6 Analysis

In this section, we analyze results to better understand performance improvements associated with the subtask block format and retrieved context in prompts.

4.6.1 Multitask Training for Subtask Blocks

To understand the contributing factors for the subtask block method's effectiveness, we examine whether the output format encodes helpful structural task information, or multitask training on individual subtasks predominantly drives performance. We conduct an additional experiment using the same subtask block output format, but without the multitask training for individual blocks. We use MRI as the source domain, because it suffers the most cross-domain performance drop. The results in Table 4.5 show a substantial drop in the model's performance in the absence of multi-task training, as compared to both the multi-task version and the baseline output format. This performance degradation may be attributed to increased decoding lengths.

4.6.2 Predictions for Multiple Anatomy Parents

In addition to differences in the anatomy parent class distribution across domains, the three examination modalities also differ in how frequently sentences with multiple anatomy entities involve multiple parent classes. As shown in Table 4.6, 57% of the sentences with multiple anatomy entities in the target domain (PET) have multiple parents, whereas the percentage is much lower for the other domains (only 12% for MRI). When using the vanilla method, models trained on a domain with few instances of multiple parents will tend to predict the same parent class for each entity, as shown by the lower frequency of prediction in the table. The use of subtask blocks together with multitask training substantially improves the model's ability to identify multiple parent types when there are multiple anatomy entities. In all domains, roughly 20% of sentences have multiple anatomy entities, so this leads to overall performance improvement.

4.6.3 Target Domain Retrieval Filtering

To reduce the search costs, we filter the unlabeled target domain data to include only sentences with anatomy terms before running retrieval with BM25. To understand the impacts on performance, we run experiments on unfiltered data, again focusing on the MRI data where domain differences are greatest. Table 4.7 shows that filtering for anatomy not only reduces costs but also gives a small improvement in results for identifying normalized categories.

4.7 Summary

In conclusion, we present generative event extraction methods for radiology findings that improve generalization under domain shifts and reduce the inference costs. By decomposing complex tasks into simpler subtask blocks and incorporating target-domain context during the inference process, our approach enables smaller models to achieve performance similar to or better than those obtained with more decoding passes, and comparable to larger models on anatomy normalization. Our methods make efficient inference for extensive clinical notes more feasible. This work offers insights into reasoning with smaller models and using context

to compensate the reduced model size.

Radiological findings' temporal changes are important for evaluating a patient. Though this work extracts temporal descriptions already written in a single note, in practice, radiologists will look at the history of radiological exams to learn the historical characteristics, compare the nuances of changes, and understand the speed of the temporal changes. The next chapter introduces longitudinal summarization approaches to reduce the reading burden and reveal the temporal changes of radiological findings.

Table 4.2: F1 scores (%) for: non-generative mSpERT [4], generative vanilla T5 models with both multi-step pipeline and one-step joint approaches, and our proposed one-step T5 model using the building block technique. All models adopt the T5-base architecture and are initialized with ClinicalT5 [5]. **Best overall** scores are in bold, and best one-step scores are underlined.

Entity	**mSpERT**	**T5-base 3-step (vanilla)**	**T5-base 2-step (vanilla)**	**T5-base 1-step (vanilla)**	**T5-base 1-step (blocks)**
			PET → PET		
Trigger	82.4	81.9	81.9	82.1	**82.6**
Anatomy Span	65.8	**67.6**	**67.6**	66.0	66.1
Anatomy Parent	61.9	64.7	**64.9**	63.3	63.5
Anatomy Child	59.6	62.1	**62.3**	59.7	60.7
			MRI → PET		
Trigger	75.6	76.6	76.6	76.4	**77.8**
Anatomy Span	59.9	60.9	60.9	59.2	**61.1**
Anatomy Parent	34.7	**48.6**	47.1	44.9	48.3
Anatomy Child	33.5	44.6	44.0	41.2	**44.8**
			CT → PET		
Trigger	75.7	76.1	76.1	74.0	**76.6**
Anatomy Span	59.7	**61.4**	**61.4**	56.3	59.8
Anatomy Parent	53.2	**55.8**	54.8	50.8	55.0
Anatomy Child	47.5	**53.3**	51.8	48.1	51.2

Table 4.3: F1 scores (%) for T5 anatomy classification models with and without contexts. Results with context involve a first pass with the 1-step T5-base building blocks method, the same as "T5-base one-step (blocks)" in Table 4.2, followed by another pass that normalizes the anatomy spans that are previously detected by the 1-step T5-base (block) model. We normalize with the model used in the last step of the 3-step (vanilla) pipeline, optionally augmented with contexts in the prompts. We also add the T5-large normalization model without context to compare with the larger-scale counterpart.

Normalization model	**T5-large**	**T5-base**	**T5-base**	**T5-base**	**T5-base**	**T5-base**
Context	**n/a**	**n/a**	**adjacent sentences**	**metadata &header**	**BM25 retrieval**	**all combined**
	PET → PET, Trigger: 82.6, Anatomy Span: 66.1					
Anatomy Parent	63.6	**63.9**	63.8	63.7	63.8	63.7
Anatomy Child	60.9	60.9	61.0	**61.1**	60.3	60.4
	MRI → PET, Trigger: 77.8, Anatomy Span: 61.1					
Anatomy Parent	51.2	50.8	52.1	51.6	**53.8**	53.5
Anatomy Child	48.6	45.4	47.1	46.6	48.3	**48.8**
	CT → PET, Trigger: 77.8, Anatomy Span: 59.8					
Anatomy Parent	54.1	54.2	55.5	55.0	55.5	**55.9**
Anatomy Child	51.2	51.2	52.2	51.6	52.6	**53.0**

Table 4.4: Average number of decoding passes per sample (indicating relative decoding time) and tokens per sample (indicating relative cost) of one-step and multi-step approaches for testing on the PET domain. The token counts per sample are the average of the sum of input and output token counts, which is used for proportionality pricing LLM usage by ChatGPT. The context method uses all context combined in another normalization step as in Table 4.3.

Method	**passes/ sample**	**tokens/ sample**
3-step (vanilla)	2.5	355
2-step (vanilla)	1.7	199
1-step (block) + context	1.7	450
1-step (block)	1	245

Table 4.5: F1 scores (%) for the cross-domain MRI-PET condition using 1-step T5-base models, comparing: vanilla output format, building block format but no multitask training, and building block format with multitask training.

Entity	**vanilla**	**blocks, no multitask**	**blocks, multitask**
Trigger	76.4	76.0	**77.8**
Anatomy	59.2	57.1	**61.1**
Parent	44.9	38.6	**48.3**
Child	41.2	36.9	**44.8**

Table 4.6: Relative frequency (%) of sentences with multiple anatomy entities that have different parents, comparing frequencies as predicted by different models to the frequencies based on gold annotations for training data. The gold relative frequency on the PET test data is 55%.

Domain	**Training**	**Vanilla**	**Blocks**
PET	57	53	**56**
MRI	12	29	**46**
CT	33	45	**52**

Table 4.7: Normalized anatomy F1 score (%) for the MRI-PET condition, comparing approaches for using target-domain context retrieved using BM25: no context, unfiltered retrieval, and filtering the retrieval corpus to anatomy informative sentences.

Entity	**no context**	**unfiltered contexts**	**filtered contexts**
Parent	50.8	52.7	**53.8**
Child	45.4	47.4	**48.3**
Trigger: 77.8, Anatomy: 61.1			

Chapter 5

LONGITUDINAL RADIOLOGY REPORT SUMMARIZATION

5.1 Longitudinal Summarization as Timeline Generation

Tracking changes in status of findings in longitudinal radiology reports is critical for understanding disease progression, treatment response, and supporting disease risk assessment. Reading longitudinal reports can capture subtle changes that may be missed in isolated reports. Individual reports often document changes in status with short phrases (e.g., "improved", "persists") [122, 123], but the longitudinal sequence of reports can provide details that make this information more useful. However, reading longitudinal records and pulling together related findings can be time consuming and contributes to cognitive overload. Artificial intelligence (AI) can support this work, both in the identification and presentation of relevant findings.

While AI tools for summarization are effective for many tasks, most work on multiple documents has involved unstructured summaries, which complicates fact-checking of statements about temporal changes [102]. For high-stakes clinical scenarios, the ability to confirm source details behind an AI statement is critical. Verma et al. [15] propose a verifiable structured summarization format in which each fact is linked to its source document and organized into human-curated topics. A pre-defined set of topics is useful for large document collections, but this is not practical for summarizing radiological findings for specific patients, where it may be useful to track multiple nodules observed in the same region of the lung, for example.

In this chapter, we introduce a new **two-dimensional timeline format for structured summarization** that makes it easy to see temporal changes. As shown in Figure 5.1 for lung-related findings, each column corresponds to a time-stamped radiology exam (as indicated in the header), such as a chest CT taken in May of a given year. Each cell in a column corresponds to a lung finding fact from that report, including details relevant to lung

cancer risk assessment (e.g., size, location, status, etc.), such as ”stable subsolid pulmonary nodule in the right upper lobe.” Each row represents a group that tracks the evolution of a particular finding over time, with a row header describing the distinguishing characteristics of that group.

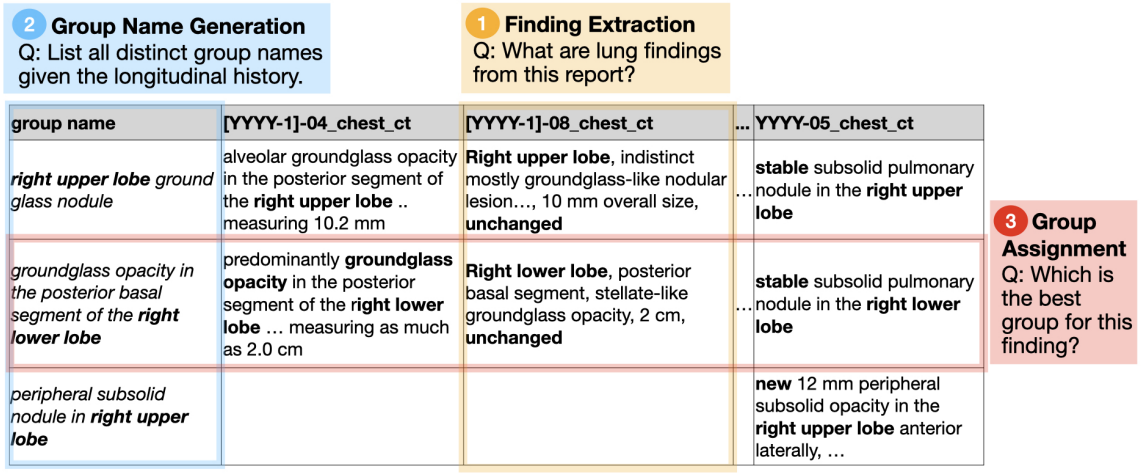

Figure 5.1: The timeline task and a three-step LLM approach. Each column corresponds to a time-stamped radiology exam (e.g., YYYY-05_chest-ct). Each cell in a column is a piece of lung finding fact (e.g., "stable subsolid pulmonary nodule in the right upper lobe") from that report including clinically important details (e.g., "stable", "subsolid"). Each row groups temporally related findings, with a row header describing the distinguishing characteristics of that group (e.g., "right upper lobe ground glass nodule").

To support research with this framework, we create **RadTimeline, an evaluation dataset** with lung-related radiological findings within longitudinal chest-related imaging reports where each patient is associated with a hand-corrected timeline. In addition, metrics are proposed for assessing automatically generated timelines in terms of finding factuality, group name quality, and finding grouping performance.

We propose a **3-step LLM-based approach to timeline generation** that: (i) extracts lung-related finding facts from individual reports (the cells in a column), (ii) generates interpretable group names (row headers) from longitudinal reports, and (iii) associates each finding with its corresponding group (row).

In experiments with RadTimeline, we show that this 3-step timeline generation approach

detects most lung findings correctly, but further work is needed to address irrelevant findings. We observe that group name quality is critical for effective grouping, particularly for weaker LLMs and embedding-based grouping strategies. The best prompting configuration reaches near-human performance in grouping findings.

5.2 Gold Timeline Dataset Curation

We sample 10 patients from a patient cohort from an existing case-control dataset created for a lung cancer surveillance project in UW Medicine,[1] where temporal changes of lung-related findings are clinically important risk factors. We ensure that each of the sampled patients having at least four chest-related radiology exams, including chest CT, chest X-ray, and abdomen CT, within a five-year window. According to the state cancer registry, 6 patients have future lung cancer diagnosis records within 3 years, and 4 patients have no diagnosis records of any type of cancer. The raw EHRs contain 65 chest-related exam reports from all patients. The number of reports per patient ranges from 4 to 12, and longitudinal report lengths measured in tokens range from 646 to 4625, with an average at 2055.

We create a timeline for each patient based on their longitudinal radiology reports, as illustrated in Figure 1. Each column represents the findings (**gold findings**) extracted from an individual radiology report, while each row captures the temporal evolution of a specific finding mentioned across multiple reports (**gold group assignments**). Each row is further assigned a **gold group name**, summarizing the key characteristics of the findings in this row. Since our focus is on tracking lung-related conditions, findings pertaining to other organs (e.g., the aorta) or descriptions of medical devices are considered irrelevant and are omitted from the timelines. We work with two medical students and one medical doctor as annotators during the timeline creation process. We first automatically generate timelines using LLMs (see Section 5.3 for details) and hand-correct LLM generated findings and group assignments based on the feedback from two medical student annotators. Finally, a third annotator reviews the gold finding groups, and hand-corrects row names of each version of gold group assignments to produce the gold group names.

[1]A subset of this cohort was previously used by Zeng et al. [124] for lung cancer prediction.

Annotation guidelines are included in Appendix B.1.

5.2.1 Gold Findings (Columns)

Factuality evaluation: We ask two medical student annotators to rate the factuality of each generated finding against its associated report, and to add any missing findings that the model fails to include. In addition, they provide notes explaining the identified errors, which are later used by an author to manually correct the generated findings as the gold-standard reference.

To reduce annotation efforts and focus on abnormalities, we manually drop absent findings, such as ”no new nodules,” and normal findings, such as “clear lungs.” The original automatically generated timelines contain 221 findings from 61 reports. After removing absent findings, we obtain 155 findings from 55 reports for the annotation tasks in factuality evaluation.

We define four categories for factuality rating: "irrelevant,” "wrong,” "partially correct," and "correct." "Irrelevant" refers to findings that are not related to lung radiological observations, typically describing findings from other organs or medical devices. "Wrong" refers to findings that contain errors regarding presence or temporal change trends. "Partial" is used for findings that contain inaccurate or incomplete details. For example, in "stable subsolid pulmonary nodule in the right upper lobe", annotators will consider it as "partial" if "stable" or "subsolid" are missing. Findings that include correct and complete clinically important details are labeled as "Correct." Table 5.1 shows the confusion matrix between two annotators' ratings. The two annotations reach a high agreement at 86%. No finding is considered wrong by both of the annotators, and the major mistake is including irrelevant findings (14%). In addition, one annotator identifies 3 missing findings from a single report. No other missing findings are reported.

Gold Finding Curation: We refine the LLM-generated findings as follows. Irrelevant or wrong findings are removed, correct findings are kept unchanged, and partially correct findings are either revised by adding missing details, or merged with another finding with the complementary details when they are both described in proximity within the same paragraph.

	Wrong	Partial	Correct	Irrelevant
Wrong	0	0	2	0
Partial	0	12	3	0
Correct	2	9	106	2
Irrelevant	1	0	2	16

Table 5.1: Confusion-matrix of factuality annotation between two annotators. Annotators agree on factuality ratings for 86% of 155 generated findings.

Additionally, we include the three missing findings identified during review. If annotators rate a finding differently, the first author resolves the disagreements based on annotators' notes to decide the action. One exception is made for one finding labeled as correct by both annotators but later found to be irrelevant during post-analysis and removed from the gold references. For four partially correct findings in which incomplete details result from complementary information spread across different parts of a radiology report, we create a new finding and add a same-report link between the new finding and the original partial finding during the subsequent group assignment curation process. This is consistent with the automatically generated findings, where finding details from different parts of reports are not always merged.

After resolving disagreements, we find that 15% of the 155 findings used in the factuality human evaluation can be dropped due to being wrong or irrelevant and 12% could be further merged with another finding in the same report. After removing and adding findings based on ratings and notes from annotators, there are 136 gold findings, or 115-116 merged findings after merging same-report linked findings grouped separately by the two annotators.

5.2.2 Gold Group Assignments (Rows)

To collect group annotations for temporally related findings, annotators assign a group ID to each finding. We provide them with automatically generated group names linked to the corresponding group IDs, displayed as the row headers in the generated timelines.

Statistic	Ann-1	Ann-2
# **Groups**	65	62
Avg. / Max size	2.1 / 8	2.2 / 7
% Non-singleton	52	50
# **Findings** (Total # 136)		
% Linked (Any / Cross)	77 / 70	77 / 71
# **Findings** (Merg.)	116	115

Table 5.2: Group annotation statistics from two annotators (Ann-1 and Ann-2) on findings in RadTimeline. The group size is based on the 136 findings. "Findings (Merg.)" are findings that are merged with same-report links for finding extraction evaluation. The % Linked refer to the percentage of findings that have links to any other finding (Any) vs. only cross-note findings (Cross) among the total 136 findings.

Annotators may add new group names and IDs if the existing list is incomplete or not representative of a group. When two group IDs convey the same meaning, annotators are instructed to select one ID and apply it consistently across all relevant findings. Revising the group names is not required at this stage. Annotators have access to the full longitudinal radiology reports while performing group ID assignment. For the seven added findings (3 missing, 4 due to missing details), group IDs have not been pre-assigned. To assign IDs to these findings, one annotator labels the seven findings in their own group assignment version, and the first author propagates these labels to another annotator's version to maintain consistency.

We collect two versions of group assignments to capture individual annotator preferences. For example, for "low lung volumes bilaterally", and "bibasilar opacities, likely atelectasis", one annotator groups them and the other separates them into different groups. The CoNLL F1 between two gold group assignments is 82, reflecting good inter-annotator agreement. Group annotation statistics are in Table 5.2.

5.2.3 Gold Group Names (Row Headers)

The third annotator who is a medical doctor reviews both versions of gold finding groups. The annotator is asked to manually correct row names of each version of gold group assignments to produce distinguishable group names, and each name should cover the shared information within the group.

5.3 Methods

We use a three-step approach to generate timelines. First, lung-related findings are identified for each report. Second, given the longitudinal report contexts, a list of group names is generated to represent distinct findings. Lastly, each finding is assigned to a group, optionally selected from already generated group names. This section describes the different configurations explored for each step, including the version used in automatic timeline generation for annotation. We evaluate three LLMs, Llama 3.1 8B instruct [125], referred to as Llama 3.1, GPT-4o [126], and GPT-OSS [127].[2]

5.3.1 Step 1: Finding Extraction

We first extract lung-related findings from each report independently using an LLM. The prompt includes the full texts from a single report, detailed instructions about generating the lung findings, a short output example, and finally a request for the final answer. The instruction asks for details that may be relevant to risk evaluation, such as size and change. Because lung finding descriptions can be lengthy, we require the output format as a Python list in order to simplify post-processing.

When generating the timelines for annotation, we use Llama 3.1 in Step 1.

We automatically remove findings that contain only absent events according to an information extraction model, RadGraph [128].[3] We use two versions of absent prediction filters. The first filters out sentences where all detected entities are absent. The second is

[2]The LLM prompts for each step are provided in Appendix B.2.

[3]The RadGraph package is available at: https://pypi.org/project/radgraph/. Specifically, we use the model version `modern-radgraph-xl`.

more aggressive, dropping predictions when the core entities (e.g. lesions) are all absent even if their descriptors (e.g. lung) are not. For the 203 Llama-predicted findings, the first filter retains 187 findings, and the second reduces these to 165, closer to the 155 findings obtained in human labeling.

5.3.2 Step 2: Group Name Generation

Instead of directly clustering the findings, we use an intermediate step to generate interpretable group name in natural language first, which are used in the later group assignment step to map findings. By comparing findings with group names, we can avoid expensive pairwise finding comparison. Group names should be specific enough so that they can separate unrelated findings, and generic names should be avoided.

The prompts include longitudinal report records of a patient, detailed instructions asking for temporally related lung finding groups, one good (specific) and one bad (generic) group name example, and the final request asking for a list of group names separated by commas.

We explore two longitudinal input contexts to represent patient records; one includes the full reports, called **full context**, and the other includes only predicted findings from all reports in Step 1, called **finding-only context**. We use finding-only context to test whether filtering content into shorter lengths improves reasoning. In both input context versions, the report contents are marked by dates and exam titles.

We use Llama 3.1 with full context in Step 2 to generate the timelines for annotation.

5.3.3 Step 3: Group Assignment

In this step, we assign each finding to a group name, so that findings with the same group name are grouped. We propose both LLM-based and more efficient embedding-based methods, each of which leverages the group names predicted in Step 2.

LLM-based Approach

We explore prompting variations including single finding (**Single**) or multiple findings (**Multiple**) per LLM run, zero-shot (**ZS**) or few-shot (**FS**), and using full longitudinal

contexts (**full context**) or **no context**.

For **Single** prompting, the LLM is prompted to assign each finding individually to the group best describing this finding among the provided group name list, or "Other" if no suitable group exists. The group name list is predicted in Step 2 for the same patient. If the prediction is 'Other', the finding's own text is used as its group name. The group names from Step 2 that are not assigned to any findings will be dropped from the final timelines. We instruct the LLM to place the selected group name in the last line of its response. During post-processing, the first group name appearing in the last line is parsed as the prediction, even if multiple group names are present.

When the generated group names are incomplete or duplicated, findings from the same group may not share consistent group names. To encourage consistent group assignment across all related findings, we experiment with **Multiple** prompting, which assigns groups to all findings of a patient in a single LLM run by inserting a tag next to each finding, following the approach of Sundar et al. [108]. Those tags serve as interpretable group names in natural language, and are parsed from angle brackets as the predicted group name for the corresponding finding. We use two variations. One does not require existing group names in the prompts (**no groups**), and the other requires them so that LLM can choose from existing group names or add new ones. This Multiple prompting is cheaper than Single prompting due to fewer runs.

We experiment with using all longitudinal radiology notes (**full context**) to provide complete cross-document information. This setup is similar to the human annotation setting where annotators have access to the full set of longitudinal reports. However, it remains unclear whether LLMs can effectively utilize such long contexts, and processing with full context is computationally expensive. Therefore, we also evaluate **no context** prompts, where group assignment relies solely on the provided group names and the finding texts.

The few-shot (**FS**) prompts are motivated by formatting errors observed when zero-shot (**ZS**) prompts are used with Llama 3.1. We manually curate two short examples without using any examples from the evaluation dataset. Each example includes the existing group names if in use, and the output in the required format.

The final prompts integrate the above variations in the following order: longitudinal

reports (omit for no context), detailed instructions (varies for Single or Multiple), two examples (omit for ZS), group names list (omit if no groups), a single or multiple findings to be grouped, and a final output request emphasizing on the output format.

We use Single ZS full-context with Llama 3.1 in Step 3 to generate timelines for annotation.

Embedder-based Approach

We also explore embedding-based methods as more efficient alternatives for LLMs in the group assignment step. We first use a non-generative embedder model to convert each finding and each group name into embeddings, then assign each finding to the group name with the highest cosine similarity from the same patient. No new groups will be added beyond the provided group names.

Our embedder backbone is E5-Mistral which is built for general domain and supports instruction integration for task adaptation. We first describe the task using a **general instruction**, "Given a radiology finding, find the group that it belongs to", which is prepended to each finding. We further infuse task-specific preference using a **task-specific instruction**. We hypothesize that finding types (e.g., "opacities", "nodules") and anatomy locations (e.g., "right upper lobe") are critical information that distinguishes lung radiological groups. Therefore, we add a prefix, "Represent the radiological findings and be aware of the type and location of the findings", to both findings and group names.

5.4 Experiment and Results

We evaluate Llama 3.1 and GPT-4o for prompting methods in all steps. Context lengths of both are 128k tokens. We consistently choose greedy decoding for Llama 3.1 and set the temperature at 0.1, top-p at 0.9 for GPT-4o, without hyperparameter tuning. The group assignment step is evaluated on gold findings.

5.4.1 Finding Extraction

We compute ROUGE-L scores at both the report level and finding level, in both cases by first scoring against each of the two annotators and then averaging the results. For

the report-level score, we represent a report as the concatenated list of all findings before automatically removing absent findings, compute the ROUGE-L score for the list, and then average ROUGE-L scores across reports. When calculating the finding-level scores, we use the merged gold findings (findings from the same report assigned to the same group) and represent them using a concatenation of the respective finding strings. Then, we automatically align each predicted finding to the gold (merged) finding that yields the maximum ROUGE-L among all the gold findings from the same report, after filtering out absent findings. The finding-level ROUGE-L score is the average over the ROUGE-L scores for all matches.

Gold findings that have no match are considered **missed**. We reversely match each gold finding to the prediction that yields the maximum ROUGE-L and report the percentage of unmatched predictions as **unnecessary**. Under an ideal alignment, unnecessary predicted findings would include wrong, irrelevant, redundant, and unfiltered absent predictions.

Table 5.3 reports the different scores for finding extraction for three LLMs, Llama 3.1, GPT-4o and GPT-OSS. Report-level ROUGE-L scores for Llama 3.1 and GPT-4o are close, and the GPT-OSS score is lower. For finding-level scoring, we provide two configurations, one with a weaker and one with a more aggressive absent finding filter. For both configurations, GPT-4o outperforms Llama 3.1 (67 versus 62). GPT-4o also has substantially lower unnecessary predictions. For examples, see Appendix B.3 Table B.9. GPT-OSS again has low ROUGE-L finding scores, but similarly low missed findings and the unnecessary rate is better than Llama 3 (but worse than GPT-4o). All LLMs have low rates of missed findings. As expected, the more aggressive filtering of absent findings leads to substantial reduction in the percent of unnecessary findings. In addition, the finding-level ROUGE-L score improves, with minimal increase in missed findings. Because of the low finding-level ROUGE-L score, GPT-OSS is not used in subsequent experiments. The ideal alignment of reference to gold findings would yield roughly 3% missing and 39% unnecessary findings for Llama 3.1 predictions, which is close to the ROUGE-L alignment, suggesting that the max ROUGE-L alignment is not a bad proxy.

Model	RL (Rpt.)	RL (Fdg.)	Miss. %	Unnec. %
Llama-3.1	67	62	3	40
GPT-OSS	61	58	3	32
GPT-4o	68	67	5	25
Llama-3.1*	-	70	3	32
GPT-OSS*	-	66	5	21
GPT-4o*	-	75	6	15

Table 5.3: Finding extraction evaluation against gold findings. ROUGE-L (RL) scores are at the report (Rpt.) and finding (Fdg.) level. Missing (Miss.) represents unmatched-gold percentage, and unnecessary (Unnec.) represents unmatched prediction percentage. The first two rows use the weaker absent finding filter; * indicates to the more aggressive filter.

5.4.2 Group Name Generation

As timeline row headers, generated group names are useful for readability, but they also impact the later group assignment step as intermediate inputs.

Similarly to the finding-level ROUGE-L scores in Section 5.4.1, we compute the group-name-level ROUGE-L scores, missing and unnecessary prediction rates, as well as the percentage of gold names that have multiple predicted versions (duplicates) under automatic alignment. We compare the generated group names against two versions of gold group names separately, and report the average.

Table 5.4 reports the scores for the generated group names from two LLMs, using either full contexts or finding-only contexts, as well as the oracle scores evaluated between the two versions of gold group names. To calculate the oracle scores, we evaluate one gold group version against the other alternatively and take the averages. When using full contexts, GPT-4o has a higher ROUGE-L than Llama 3.1 (48 versus 38), but also a higher unmatched gold percentage (30% versus 25%). The Llama 3.1 model has nearly 60% unmatched predictions, suggesting considerably more unnecessary group names, likely due to irrelevant

and duplicative predictions. When switching to finding-only contexts, the duplication issues in Llama 3.1 (from 32% to 42%) and unnecessary predictions in GPT-4o (from 37% to 46%) become more severe. In contrast, the missing percentage of GPT-4o decreases from 30% to 13%. These results suggest that using itemized findings as contexts produces a longer list of group names. Examples are given in Appendix B.3 Table B.10 and B.11.

Method	RL	Miss. %	Dup. %	Unnec. %
Llama 3.1 (Full)	38	25	32	60
Llama 3.1 (Fdg.)	42	27	42	58
GPT-4o (Full)	48	30	28	37
GPT-4o (Fdg.)	52	13	31	46
Oracle	87	9	11	9

Table 5.4: Evaluation results for group name generation. The groups are generated using full or finding-only (Fdg.) contexts.

5.4.3 Group Assignment

We evaluate group assignment methods on gold findings, in order to directly compare with their associated gold group assignments. We consider findings that share the same group name and come from the same patient as a group, and report CoNLL F1 scores as in CDCR tasks. Given that we have two versions of gold group assignments, we calculate CoNLL F1 against each version and report the averages. We also report oracle results, assuming that gold group names are known and used as intermediate inputs, and further average the scores when evaluating against two gold group name versions separately.

Comparing Prompting and Embedder Methods without Contexts

Table 5.5 compares prompting methods and embedding-based methods when using no context in Step 3. We evaluate these methods when using generated group names, gold group names

Group names	Single ZS Llama	Single FS Llama	Multi. ZS Llama	Multi. FS Llama	Single ZS GPT	Single FS GPT	Multi. ZS GPT	Multi. FS GPT	Embed. Gen.	Embed. Spec.
Llama-3.1	66	74	77	78	77	73	78	77	68	70
GPT-4o	75	79	77	81	82	80	84	82	65	69
Oracle	73	78	86	84	80	84	85	83	78	83
No groups	n/a	n/a	72	68	n/a	n/a	70	72	n/a	n/a

Table 5.5: Group assignment performance of prompting and embedding-based methods with no context in CoNLL F1 scores. We use LLM models (Llama 3.1, GPT-4o) and an embedder model (e5-mistral-7b-instruct). "Multi." is short for Multiple prompting. "Embed. Gen." and "Embed. Spec." represent embedding-based methods using the general or task-specific instructions.

(oracle), or no group names. All generated group names are obtained from Step 2 using full contexts.

The choice of intermediate group names is crucial, and GPT-4o approaches reach near-human performance on gold findings. When shifting group names generated by Llama 3.1 to GPT-4o, all prompting methods improve, except that Multiple ZS prompt with Llama 3.1 is unchanged. The Single FS prompt with Llama 3.1 improves from 74 to 79 in CoNLL F1, and with GPT-4o it improves from 73 to 80. GPT-4o prompting methods with GPT-4o group names are all close to the inter-annotator performance at 82 CoNLL F1, and the best one, using Multiple ZS, reaches 84. Embedding-based methods do not show improvement with GPT-4o-generated names. Under oracle scenarios, prompting methods show mixed results compared to using automatically generated group names, indicating that automatic names are not necessarily worse than human curated ones. In contrast, embedding-based methods significantly improve with oracle names, and the best embedding-based method, which uses task-specific instructions, achieves near-human performance at 84 CoNLL F1, indicating their potential as efficient substitutes to LLMs when group name generation is further improved. For the Multiple prompt version, we observe a

considerable performance drop to around 70 after removing group names, suggesting the importance of including the group name generation step.

Domain knowledge helps Embedding-based methods. We compare the task-specific instruction and the general instruction for embedding-based methods. It shows that adding domain knowledge to instructions that emphasize the finding types and anatomy locations consistently boost results under all group name settings (from 78 to 83 with oracle names). The gap to LLM-based methods is narrower but still persists. More task adaptation efforts on embedding-based methods are needed.

Few-shot prompting only helps Llama 3.1. We observe that few-shot examples help Llama 3.1 in following a specified format. However, for GPT-4o, few-shot examples do not seem necessary.

Impact of Full Contexts

	Single ZS Llama	Single FS Llama	Multi ZS Llama	Multi FS Llama	Single ZS GPT	Multi FS GPT
Annot. 1						
No Ctx.	67	74	79	78	81	82
Full Ctx.	56	70	76	78	77	81
Annot. 2						
No Ctx.	65	75	75	77	82	85
Full Ctx.	60	76	82	78	86	90

Table 5.6: Impact of full contexts in group assignment prompting methods. Evaluated in CoNLL-F1 against two gold group assignments separately. Each configuration uses either Llama 3.1 for all steps or uses GPT-4o for all. All experiments use full contexts by default in group name generation. "Ctx." is short for "Context".

Table 5.6 shows the impact of adding full contexts for group assignment prompts,

evaluated against two versions of gold group assignments separately. With the exception of Llama 3.1 Single ZS prompting, where full context hurts using both gold annotations, full context improves performance when measured against gold group assignments from annotator 2 but not annotator 1. These results suggest annotator differences in leveraging long-contexts, which is worth future investigation.

Group names	Single ZS	Single FS	Multi. ZS	Multi. FS
Llama (Full)	66	74	77	78
Llama (Fdg.)	66	70	71	68
GPT (Full)	82	80	84	82
GPT (Fdg.)	73	74	74	73

Table 5.7: Impact of input contexts for group name generation on the final group assignment performance in CoNLL-F1. All experiments use the same LLM in all steps and no contexts in group assignment. Group names are generated using full or finding-only ("Fdg.") contexts.

In Table 5.7, we further evaluate the impact of different group name generation prompts on group assignment. We observe that switching from full contexts to finding-only contexts for group name generation hurts the final group assignment performance, indicating that the additional context is useful.

5.5 Conclusion

We propose a timeline generation task for structured summarization on longitudinal radiology reports. This timeline format groups temporally related findings for straightforward comparison and facilitates fact-checking by denoting the associated report for each finding. We create a timeline dataset, RadTimeline, to evaluate timeline generation. We propose a three-step LLM approach that can achieve near-human performance in grouping gold findings.

This work can be extended to a broader clinical longitudinal records to capture temporal nuances in clinical narratives, such as for symptom progression, medicine history, and

treatment trajectories.

This work has several limitations. Our method for evaluating grouping of findings requires that the findings have gold group assignments, which are only available for the gold findings. One solution would be to assign "silver" labels to the predicted findings by aligning them to gold findings, as in the finding assessment approach. Of course, this strategy should be validated further through human evaluation of the fully automatic timeline. Another limitation is that the evaluation dataset is relatively small, due to the costs associated with clinical excerpts and the cognitive effort required to understand longitudinal reports. Meanwhile, this evaluation dataset gold annotations may share a similar language style to Llama 3.1, because the dataset is hand-corrected from generated timelines produced by Llama 3.1 based on expert feedback. This could lead to bias in ROUGE-L score. Further human evaluation on different LLMs is recommended.

Chapter 6

LUNG CANCER RISK PREDICTION USING EXTRACTED RISK FACTORS

Lung cancer is the top cause of cancer-related deaths [129], but early detection and treatment can improve survival rates. Selecting high-risk individuals of lung cancer for low-dose CT (LDCT) screenings can reduce mortality by over 20% [130, 131]. Unstructured EHR notes contain rich lung cancer risk factor information that can complement structured EHRs for risk modeling. In this work, we implement lung cancer risk prediction models that leverage both structured data and information extracted from unstructured notes, showing the benefit of unstructured data in two different models. We also highlight challenges for language models in risk prediction arising from the inclusion of EHR information that is not pertinent to lung cancer risk, as well as from leveraging longer time periods of patient history.

6.1 Lung Cancer Risk Prediction Background

Rule-based criteria based on age and smoking history are used to enroll people who have smoked at some point in their life (henceforth referred to as ever-smoker) in lung cancer screening trials [130, 131], and to guide preventive recommendations in the United States Preventive Services Task Force (USPSTF) guideline.[1] Bach et al. [132] find individual risk varies in the ever-smoking group, and proposes the Bach model for predicting 10-year lung cancer risk based on age, gender, asbestos exposure history, smoking years, and quit years. The LLP model [133] incorporates additional risk factors, including family history of cancer, disease history of pneumonia, and other cancers, for 5-year risk prediction. PLCOm2012 [134] predicts a 6-year risk with additional clinical information, including body mass index, chronic obstructive pulmonary disease (COPD), and previous chest radiography, showing

[1]Adults aged 50 to 80 years who have a 20 pack-year smoking history and currently smoke or have quit within the past 15 years are recommended for annual lung cancer screenings.

that their individualized risk scores have greater accuracy in determining screening eligibility than the rule-based clinical trial eligibility criteria.

After the initial screening, the CT radiological findings, especially nodule-related information, can further stratify high-risk patients to guide decisions in follow-up screenings or diagnostic tests. In addition to pre-LDCT risk factors, the PanCan model [135] incorporate nodule features to predict 2-year malignancy risk for a pulmonary nodule after the initial LDCT exams, predictive features including size, spiculation shape, upper lobe location, attenuation (solid, part-solid or others), and count. To incorporate prior CT results if available, Tammemagi et al. [136] combine three annual rounds binary outcomes with the pre-LDCT PLCOm2012 scores, showing imprement in predicting lung cancer risk of 1 to 4 years. New nodules, even if small, are more likely to become malignant [137]. The speed of nodule growth, measured by the volume doubling time, can stratify lung cancer risk among patients with intermediate-sized nodules (5-10mm) [138]. Huang et al. [139] use non-nodule risk factors and nodule features from two rounds of screening to accurately predict 3-year lung cancer risks (AUC 0.899). Beyond extracted nodule features, deep learning image models [140, 141] predict lung cancer risk using chest CT images only, assuming CT images reflect other risk factors such as smoking status [140, 141] While those approaches [140, 141] rely on expert-annotated malignancy localizations, LungEvaty [142] uses whole-lung inputs to improve long-term risk prediction, suggesting the importance of modeling the entire lung instead of isolated regions.

The data sources of the majority of above risk prediction models are clinical trials. Well-known examples are the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Randomized Controlled Trial [143], the National Lung Screening Trial (NLST) [130], and the Dutch–Belgian lung-cancer screening trial (NELSON) [131]. PLCO [143], conducted with 78,216 participants aged 55 to 74 from 1993 to 2001 in the US, assesses whether annual chest X-ray screenings could reduce lung cancer mortality. However, it shows no significant mortality reduction compared to usual care. It collects non-imaging lung cancer risk factors, such as demographic information and smoking history, which is used to develop pre-LDCT risk prediction models [134, 144]. NLST, enrolling 54,000 55 to 74 year-old ever-smokers from 2002 to 2004, demonstrates a 20% reduction in lung cancer

mortality with three annual screenings of chest LDCT. Compared with PLCO, NLST focuses on ever-smokers and uses chest LDCT instead of chest X-ray. This supports risk prediction methods that consider LDCT findings [140, 141, 139, 136, 145]. The Nelson study recruits heavy smokers aged 50 to 75, including 13,195 men and 2,594 women from the Netherlands and Belgium. It differs from NLST by using nodule volume instead of diameter for follow-up decisions and increasing the intervals for later screenings. This dataset supports models that access nodule malignancy probability [138, 137].

Compared to using clinical trial data, using clinical routine EHR data can automate identification of high-risk patients to reduce physician burden [146], while also can leverage cheaply obtained EHR signals, and be more inclusive for younger individuals and non-smokers beyond [147]. Chandran et al. [146] predict a 3-year lung cancer risk for adults aged 45 to 65 years using structured EHR data, achieving an AUC of 0.76, with age, smoking, race, ethnicity, and chronic obstructive pulmonary disease as the top predictors. Lung cancer risk signals can also come from the longitudinal EHR nature and the unstructured data. Sequential medical codes provide early predictive signals for lung cancer risk [148, 149, 150]. To model longitudinal records, Yeh et al. [149] used CNN layers to capture patterns in diagnosis and medication codes. Temporal representations, such as visit-level positional embeddings and explicit time-gap tokens, are also found helpful when using transformer-based models [148]. Unstructured longitudinal EHRs can be further incorporated as textual embeddings [151] or extracted medical concepts (CUIs) [152] to augment missing information in structured EHRs.

We present a retrospective study for three-year lung cancer risk prediction, including patients aged over 40 years who were registered in a medical system and required a chest-related X-ray or CT examination. Individuals have varying smoking statuses and diverse demographic backgrounds, as recorded in the EHR system. Our work uses various clinical risk factors and leverages time-series EHRs. Specifically, we use readily accessible structured data, including demographic information and smoking status, along with patient note time series that document patients' radiology exam findings, smoking status, and disease history. Since CT scans are costly, we also include more accessible information from chest X-rays.

Rather than directly modeling sequences of medical codes, we use LM-based IE methods, including some approaches described in previous chapters, to extract key risk factors and generate a readable summary that captures major lung cancer risk factors, then use it for risk prediction. Given the radiological nature of the cohort, we focus on longitudinal representations of radiological findings, and represent other risk factors as either a category or a number.

6.2 Risk Prediction Tasks and Cohort Description

6.2.1 Task definition

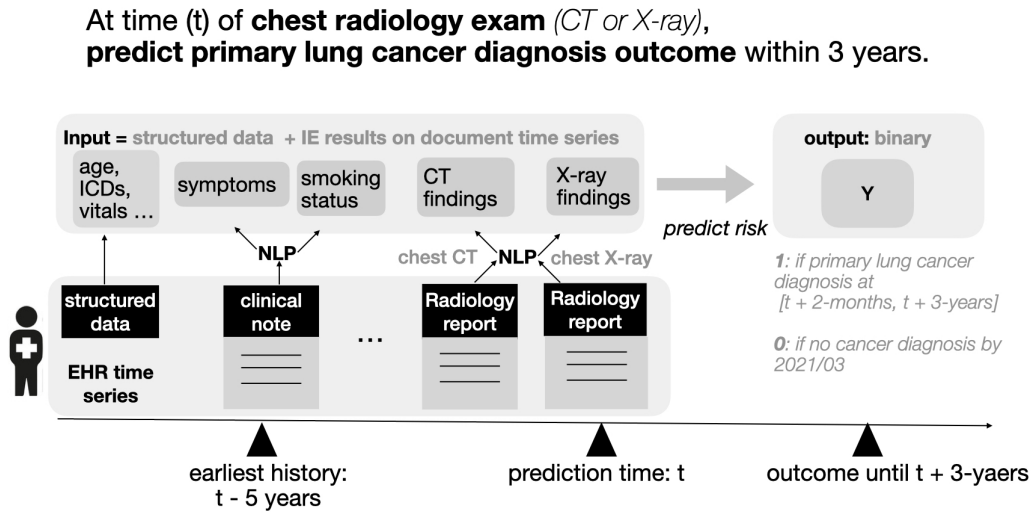

Figure 6.1: Lung cancer prediction task design. Our dataset supports predicting lung cancer within risk periods of 3 years.

As shown in Figure 6.1, the risk prediction task predicts the binary outcome of whether a patient without prior cancer diagnosis of any type will develop primary lung cancer within three years. The risk is predicted at the time of a chest-related radiology exam of certain types that reflects lung abnormalities, including chest CT, chest X-ray, and abdomen CT. We

refer to this time point as the **prediction time (t)**. A case has a positive binary outcome label, while a control has a negative one. A case corresponds to a patient and a prediction time (t), such that a diagnosis of primary lung cancer occurs within the time window [t + 2 months, t + 3 years]. If a patient undergoes multiple radiology exams within the three years prior to diagnosis, multiple cases can be associated to this patient. This 2-month prior diagnosis window ensures that we predict on pre-diagnosis information when diagnosis outcomes are not yet known, and pushes our methods to predict risk earlier than clinical diagnosis observed in practice. A control corresponds to a patient and a prediction time (t), where the patient has no cancer diagnosis of any type according to the Washington State Cancer Registry, and the prediction time coincides with one of the patient's chest-related exams. We predict the risk for each case or control using their EHRs in a 5-year **history window** before the prediction-time radiology note.

The risk prediction model outputs a risk score for each patient, with positive cases identified when the score exceeds a given threshold. Models are evaluated by sweeping over the threshold and computing the area under the ROC Curve (AUC) and the area under the Precision-Recall (PR) Curve, i.e., average precision (AP).

6.2.2 Sampling Cases

For cases, we first select eligible case patients with primary lung cancer diagnosis records, then sample prediction times for case patients when an eligible radiology exam is taken within 3 years prior to the diagnosis.

The cohort sampling is illustrated in Figure 6.2. Each **eligible case patient** must have a primary lung cancer diagnosis record with a valid diagnosis date (T) in the Washington State cancer registry, which covers the period from January 2007 to March 2022. The case patient's first recorded cancer diagnosis must be lung cancer. To further ensure identification of primary lung cancer, patients with any cancer diagnosis record that has a missing date are excluded. To identify lung cancer diagnoses, we filter the cancer registry database for records where the 'CANCERSITE' value equals 'Lung and Bronchus'.

Each **eligible case** corresponds to a prediction time for an eligible case patient and

is linked to an eligible radiology exam. The eligible prediction-time exam must be either chest CT, abdomen CT, or chest X-ray, and be completed at least 2 months before and up to 3 years before the diagnosis date (T). Patients must be 40 years old or above at the prediction time, given that young lung cancer patients are rare. For the test set, we require that patients have at least 120 days of encounter history within the five-year EHR period prior to the prediction time (t), ensuring they were active before diagnosis and have sufficient historical data as model inputs. In addition, for the test set, the case's prediction time must have been completed before March 2019 to guarantee a three-year follow-up period, as our cancer registry records end in March 2022. Due to the small number of eligible cases, those requirements are relaxed for training data.

6.2.3 Sampling Controls

Each **eligibility control patient** must have no cancer diagnosis records for any cancer type in the entire cancer registry. Each **eligible control** corresponds to a prediction time for an eligible control patient and is linked to an eligible radiology exam. The eligible prediction-time exam must be either chest CT, abdomen CT, or chest X-ray, and the patient must be at least 40 years old at that time. For the test set, as for cases, controls are required to have encounter histories longer than 120 days prior to the prediction time (t), and have the prediction-time radiology notes completed before March 2019/03 to allow a three-year follow-up period.

6.2.4 Matching Cases and Controls

For each case, we choose 10 controls matched on the note completion time, the exam type, and whether the patient has more than 120 days of encounter history within the five-year window prior to the prediction time. The note completion time is matched if two note completion dates overlap in a three-month range. The note type is matched if the two note types are identical, within the categories of chest CT, abdomen CT, or chest X-ray. We match the encounter history length as a binary condition (at least 120 days vs. shorter than 120 days) for the training data, and require test data all have at least 120 days of encounter

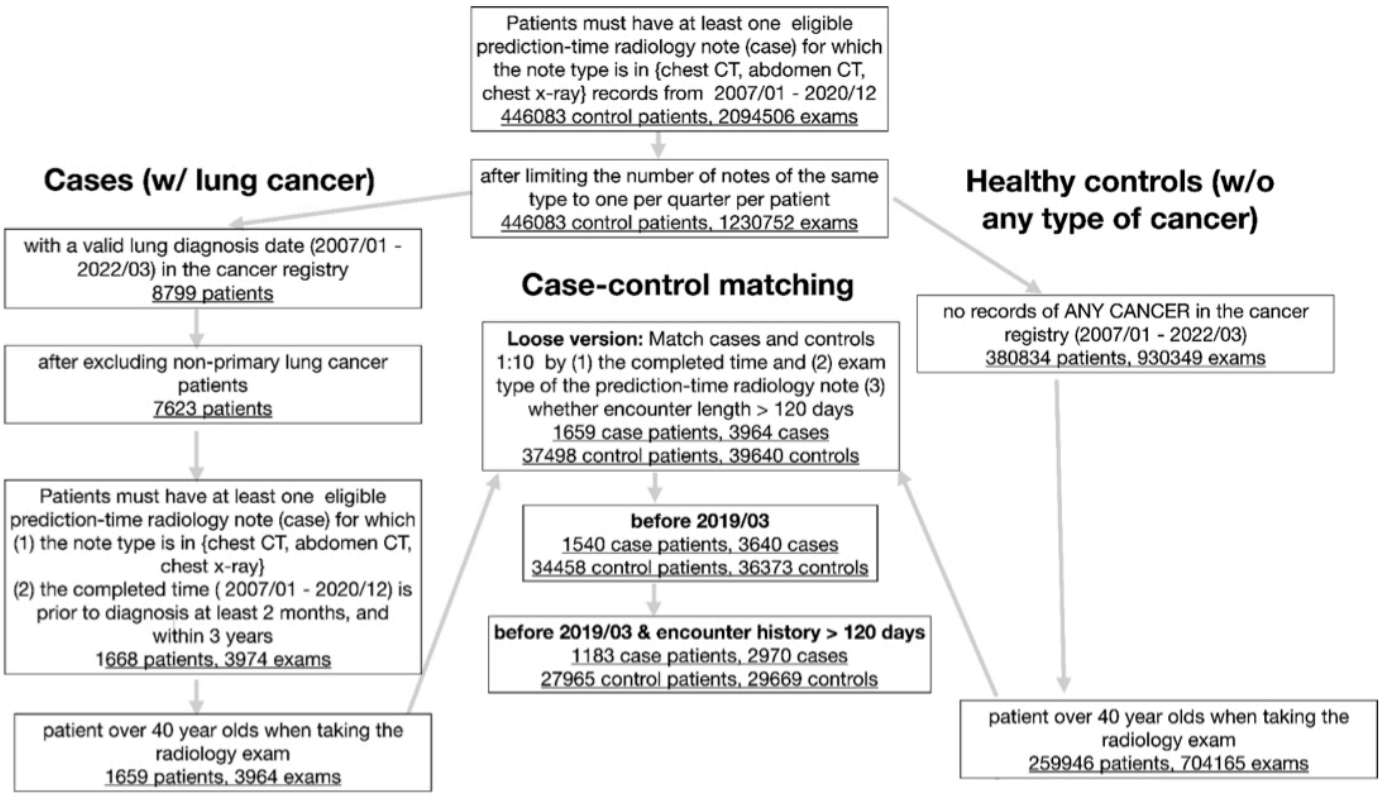

Figure 6.2: The lung cancer case-control cohort creation.

history. This strategy includes more cases in the training data compared to excluding those with short encounter histories. We ensure that each control matches only one case. When multiple controls share the same control patient, they are all matched with cases from a single case patient, in order to avoid data contamination when splitting case patients and their matched controls into training and test sets.

6.2.5 Dataset Description

Group	**Number of Patients**	**Number of Cases/Controls**
Case	1,659	3,964
Control	37,498	39,640

Table 6.1: Case and control patient counts in total including training and test sets

After cohort creation, our dataset includes 3964 cases and ten times as many controls, with dataset size details provided in Table 6.1. When splitting the training and test data, we first perform the split at the patient level, with 30% of the patients included in the test set and choosing only cases or controls with at least 120 days of encounter history before the the prediction time (t).

For the prediction input of each patient, in addition to the prediction-time radiology report, we also use additional structured and unstructured data from patient EHRs within a five-year period-history up to the current prediction time (t). The structured data include demographic information, diagnosis codes, and social determinant of health records. The unstructured data are narrative notes, including 1,126,451 clinical notes and 488,541 radiology reports, of which 232,804 are chest-related (chest X-ray, chest CT, and abdomen CT).

6.3 Methods

The input features are risk factors from EHRs, and the label is a binary diagnosis outcome. The input features can be numerical, categorical or textual. The demographic features (age, race/ethnicity, gender) are completely from structured EHRs with high data completeness.

Non-demographic categorical risk factors, including smoking status, histories of COPD, emphysema, or abnormal lung radiological findings, have structured EHRs for some patients, and we further extract them from unstructured texts using LLM-based IE methods for higher data completeness. To represent longitudinal lung radiological findings, we further use a textual feature to describe lung findings' histories and their temporal changes, obtained exclusively via IE from unstructured EHRs.

The following subsections describe the predictors, detailing each input feature in terms of its EHR source, the IE methods if using unstructured data, and its representation in the model when not used in raw format.

6.3.1 Predictors

To accommodate different types of features, we use two predictors to predict the lung cancer risk, including XGBoost [153], which is suitable for numerical and categorical features, and DistilBERT [154], which takes text inputs and can incorporating all features after converting them into texts.

XGBoost is a widely used predictor for tasks involving categorical features. We use XGBoost for the baseline with only structured data, and with the same categorical features updated with additional information extracted from unstructured notes. We train XGBoost with the logistic objective to predict the binary outcome.

To further include textual features, we combine word-based representation of structured data and findings from unstructured data. using a textual template, and feed the final text into a DistilBERT model. Compared to XGBoost, the LM approach has the advantage that it can incorporate free-form text features. We use the DistilBERT-base model with maximum input length 512 tokens to predict a binary diagnosis outcome. The model is trained using mean squared error loss as in regression tasks on the binary labels, as preliminary experiments show better performance than using cross-entropy loss. To prepare the text inputs for DistilBERT, we use an input template: "This patient is a <age-in-words> year old <ethnicity> <gender> <smoking phrase>, <COPD/emphysema phrase> <abnormal-lung phrase> <longitudinal radiological findings>." We truncate the input text at the end if

its length exceeds 512 tokens. This format provides a concise descriptive summary of the patient's pertinent risk factors related to lung cancer, inspired by clinical vignettes used for studying clinicians' decision-making [155]. This input is potentially useful for a clinician's quick review as a summarization of the raw EHRs for interpretability.

6.3.2 Demographic Information

The demographic information includes age, gender, and race/ethnicity, stored as structured EHRs as numbers or categories. Ages are numbers reported by patients. Other types of demographic information are categorical and have high data completeness. We use gender categories, including "Female", "Male", and "Other". Race/ethnicity categories include "White", "Black or African American", "Asian", "American Indian or Alaska Native" and "Other". "Other" includes missing records and low-frequency categories. To convert them into texts when using DistilBERT, we verbalize age into words, for example, "62" is converted into "sixty-two", and verbalize categories using their textual labels.

6.3.3 Categorical Risk Factors

6.3.3.1 Smoking Status

Smoking status is characterized with four categories: current smoker, quit smoker, never smoker, and unspecified. Smoking status may be logged at different times in structured data. We define valid smoker categories as current, quit, and never. Missing values and low-frequency categories (e.g., passive smoker) are grouped together treated as invalid. When multiple records are available, we use the most recent valid value.

When structured data does not contain a valid smoking status, we use an LLM to extract the smoking status category from the patient's clinical note history. We first filter smoking-related sentences from clinical notes, ordered chronologically from oldest to most recent, then we ask an LLM (Llama-3-8B-Instruct) to determine the smoking status among "current", "quit", "never", and "unknown" based on the selected sentences, using the prompt shown in Table C.1. To filter smoking-related sentences, we retain sentences that are proximal to smoking smoking-related keywords ("smok","cigar","tobacco","nicotine") in the clinical

notes, and then apply a social determinant event extraction tool [4] to select sentences containing tobacco events. To parse results from an LLM response, we search from the last sentence to the earlier sentence until we find one valid answer option.

When using XGBoost to predict lung cancer risk, smoking status is represented with 3 binary indicators for current, quit, and never, only one of which can be active. All indicators are 0 if the status is unspecified. When using DistilBERT, the smoking status is verbalized into <smoking phrase> as "current smoker" or "never smoker" or the empty string. The "quit smoker" variable was not useful in XGBoost experiments, so it was not considered in verbalization rules.

6.3.3.2 COPD

The history of COPD, which includes emphysema and chronic bronchitis, increases the risk of lung cancer [134]. We represent a patient's COPD history with two binary indicators: if they have COPD (has-COPD) or specifically emphysema (has-emphysema). With the structured data, we set the indicator positive if the patient has any diagnosis codes of International Classification of Diseases (ICD) that are related to COPD or emphysema specifically.

Since ICD codes may be missing, we further extract COPD history from unstructured EHRs, so that the final indicator is set to true if identified in either structured or unstructured source. We ask an LLM (Llama-3-8B-Instruct) if the patient has a history of COPD, emphysema, or bronchitis based on selected sentences from the patient's clinical notes and radiology reports. The prompt is shown in Table C.2. Considering the challenge of long contexts for LLM, we select sentences from both clinical notes and radiology reports instead of using their full texts. For clinical notes, we include the last five sentences that contain any of the keywords "copd," "emphysema," or "bronchitis." For radiology reports, we apply a radiological findings event extraction model [123] to full report texts to select sentences containing present medical problems or indications. During post-processing of the LLM response, the has-COPD indicator is also set to true when either chronic bronchitis or emphysema is detected.

For XGBoost, we simply use the two binary indicators. For DistilBERT, these are

verbalized as “has COPD” if only COPD is true, and “has COPD and emphysema” if emphysema is true. These phrases are then inserted into the input template as <COPD/emphysema phrase>.

6.3.3.3 Abnormal Lung Radiological Findings

In structured data, abnormal lung findings can be recorded using ICD codes. We use a binary indicator (has-lung-abnormality) to show if the patient has any abnormal lung findings. We search patient ICD codes for ICD-10-CM,R91.1, ICD-9-CM,793.1, ICD-9-CM,793.11, and set has-lung-abnormality to True if found.

Radiology reports also contain observed radiological abnormalities. In particular, CT exam reports can detect lung lesions, even small ones, and often include important lesion size details, where a larger size is associated a higher malignant risk. We ask an LLM (Llama-3-8B-Instruct) for the most recent size measurements of any pulmonary nodules, solid nodules, part-solid nodules, and non-solid-or-GGN nodules, based on selected sentences from CT reports. The prompt is in Table C.3. We select CT report sentences that contain lung lesion events detected by a radiological event extraction model [123], where the trigger type is "Lesion" and the normalized anatomy label is "Lung". In the post-processing of the LLM response, we set has-lung-abnormality to true if any of the solid, part-solid, and non-solid-or-GGN nodule is larger than 6 mm. We select this threshold because small nodules are often not concerning for lung cancer risk.

Similar to COPD, we set the final binary indicator is true if it is true in either the structured or unstructured data. XGBoost uses it directly as a binary feature. DistilBERT converts the binary indicator into a phrase <abnormal-lung phrase>, being ’has abnormal lung lesions’ if true, and an empty string otherwise.

6.3.4 Longitudinal Radiological Finding Description

With the DistilBERT model, we experiment with including textual descriptions of longitudinal lung findings from series of chest-related radiology reports. This can be a summary paragraph from an LLM, a sequence of lung finding information extracted from a series of reports using

an event extraction model or an LLM, or a reordered sequence of extracted lung findings after grouping.

LLM Paragraph Summarization

We provide the LLM with chest-related reports in chronological order, from the oldest to the most recent. The task is to summarize the lung findings with important details, such as location, size, shape, density, and the temporal changes. The prompt (as shown in Table C.4) requires the LLM to first list all the evidence related to lung findings, then summarize for each finding, and finally conclude with a concise paragraph. We use the last paragraph of the LLM response as the textual longitudinal-radiology-finding feature.

Event-based Finding Sequence - Ordered by Time

We apply the radiological finding event extraction model [123] to chest-related radiology exam reports that can detect lung abnormalities, including chest CT, chest XRay and abdomen CT.

We identify lung-related events by selecting extracted events with a trigger type of Lesion or Medical Problem, an assertion type of Present, and a normalized anatomy label of lung or pleura membrane. We represent each event as a string including the trigger span and associated argument spans that describe finding details. For the final longitudinal-radiology-finding feature, we concatenate all events in reverse chronological order, from the most recent to the oldest, so that most recent records, which are more impactful, can be preserved when the risk prediction input exceeds the maximum input length.

LLM Finding Sequence - Ordered by Time

We use the outputs from the timeline output from Section 5, and use the findings grouped by columns in chronological order as in the event sequence. We use Llama 3.1 instead of GPT-4o in order to efficiently process on all reports on a HIPAA compliant server.

note count	radiology note		chest CT		chest XRay		abdomen CT		clinical note	
	case	control	case	control	case	control	case	control	case	control
0	0	0	0.38	0.52	0.13	0.15	0.66	0.7	0.12	0.11
1	0.07	0.1	0.27	0.27	0.23	0.33	0.22	0.2	0.03	0.04
2	0.07	0.09	0.16	0.09	0.14	0.16	0.06	0.05	0.04	0.04
3-5	0.12	0.15	0.1	0.07	0.17	0.15	0.05	0.03	0.07	0.08
5-10	0.24	0.26	0.07	0.04	0.18	0.12	0.02	0.01	0.11	0.14
10-20	0.26	0.23	0.01	0.01	0.09	0.06	0	0	0.12	0.16
>= 20	0.24	0.18	0	0	0.05	0.03	0	0	0.49	0.43
total incidence	3964	39640								

Table 6.2: Distribution of note count per example of cases and controls in a 5-year history window. The numbers are relative frequency over all cases examples, or control examples.

LLM Finding Sequence - Ordered by Group

We use the Llama 3.1 outputs from the timeline output from Chapter 5 , and use the findings grouped by rows to order findings as groups. We include the generated group names in front of the grouped findings.

6.4 Experiments and Results

We first will show how unstructured data benefits data completeness. Then we show the impact of those features.

6.4.1 Data Distribution Characteristics

Table 6.2 shows the completeness of unstructured data including radiological reports and clinical notes. All patients have radiology reports, but chest CT reports, which are critical for lung cancer risk assessment, are not available in 38% of cases, and 52% of controls. Table 6.3 shows structured EHR distribution for demographic information.

Distribution	case	control
Age-at-Report		
([40, 50]	0.04	0.23
(50, 55]	0.06	0.13
(55, 60]	0.12	0.14
(60, 65]	0.15	0.13
(65, 70]	0.18	0.11
(70, 80]	0.26	0.14
(80,)	0.18	0.12
EthnicHeritage		
American Indian or Alaska Native	0.01	0.02
Asian	0.06	0.08
Black or African American	0.08	0.10
White	0.74	0.70
Declined to Answer*	0.00	0.01
Native Hawaiian or Other Pacific Islander *	0.01	0.01
Unavailable or Unknown *	0.01	0.02
Missing *	0.09	0.06
Gender		
Female	0.48	0.48
Male	0.52	0.52
Unspecified	0	0

Table 6.3: Structured EHR distribution of cases and controls in relative frequency. Ethnicity categories (*) are merged as an Other category

positive binary pct	case	control	case	control
smoking	structured		merge with LLM	
current-smoker	0.18	0.07	0.22	0.11
never-smoker	0.08	0.21	0.09	0.28
quit-smoker	0.23	0.15	0.29	0.21
COPD				
has-COPD	0.48	0.23	0.71	0.45
has-emphysema	0.19	0.04	0.41	0.13
abnormal lung finding				
has-lung-abnormality	0.28	0.12	0.40	0.16

Table 6.4: Binary risk factor feature distribution of cases and controls in relative frequency when using only structured EHRs versus combined with unstructured data.

6.4.2 Risk Factor Completeness

We show the change of feature distribution after augmenting risk factors using the unstructured data in Table 6.4. For the important risk factor, smoking status, less than 50% examples have valid structured Tobacco EHR records for both cases and controls. By merging with unstructured data using an LLM, the never-smoker feature is boosted by 33% in controls, and only 13% in cases. The current-smoker feature is boosted from 22% in cases, and 57% in controls. This indicates more missing smoking information in controls. When the positive has-emphysema features are combined with unstructured data using the LLM, the number doubles for cases and triples for controls. Positive abnormal lung feature is increased by 43% in cases and 33% in controls when augmenting with unstructured data using an LLM.

features	AUC	AP	AUC	AP
	structured data only		merge with LLM	
XGBoost demographic	0.664	0.148		
+ current-smoker	0.708	0.194	0.714	0.206
+ never-smoker	0.726	0.192	0.742	0.206
+ quit-smoker	0.665	0.153	0.669	0.159
+ has-COPD	0.708	0.182	0.704	0.180
+ has-emphysema	0.690	0.182	0.713	0.204
+ has-lung-abnormality	0.688	0.187	0.714	0.218
XGBoost (selected factors)	0.766	0.258	0.792	0.306
DistilBERT (selected factors)	0.783	0.260	0.796	0.305

Table 6.5: Impact of individual categorical risk factors using an XGBoost model and comparison of XGBoost to DistilBERT for the set of all categorical features except quit-smoker.

6.4.3 Impact of Categorical Risk Factors

Using XGBoost, we investigate the impact of individual categorical features by adding them to demographic features, as shown in Table 6.5. All structured features improve significantly from the demographic-feature baseline, except quit-smoker. Merging with unstructured data improve all features except current-smoker,quit-smoker and has-COPD. Never-smoker feature is the strongest structured feature based on AUC, achieving 0.726 AUC compared with 0.664 AUC for the demographic baseline, and unstructured data further improves the AUC to 0.742. The has-lung-abnormality benefits most from the LLM information, particularly for AP, improving the 0.148 demographic baseline to 0.218.

Table 6.5 also shows results for prediction using all categorical features except quit-smoker. We compare the categorical features from only structured data with the version merging with LLM-extracted features from unstructured data. When using only structured data,

Methods	AUC	AP	# Input tokens		
			Median	95%	Max
Demographic & Categorical	0.796	0.305			
Event sequence (time order)	0.837	0.365	62	494	4659
LLM summary	0.816	0.352	118	197	4889
LLM finding list (time order)	0.825	0.366	125	1020	9508
LLM finding list (group order)	0.816	0.325	568	2075	96995

Table 6.6: Impact of adding longitudinal radiological finding features from event extraction, LLM (Llama 3.1) summary, time-ordered LLM finding list, or reordered LLM finding list, using DistilBERT for risk prediction. To indicate potential truncation effects, we also report the input sequence lengths (in tokens), including the median, 95th percentile, and maximum. For reference, DistilBERT input length is 512 tokens.

DistilBERT achieves better AUC performance than the XGBoost model: AUC of 0.765 from DistilBERT vs. 0.745 from XGBoost. When merging with LLM extracted risk factors, both models give similar performance. Compared to using only structured data, incoporating unstructured data with LLM improves AP scores from 0.26 to 0.31, and slight gain in AUC from 0.78 to 0.80, based on DistilBERT results.

6.4.4 Impact of Longitudinal Risk Factors

In Table 6.6, we compare different longitudinal lung radiological abnormality features, when adding to the best model from the experiment with categorical risk factor features. The event sequence in time order works the best, followed by the time-ordered LLM extracted findings, then the LLM paragraph summary. LLM-extracted findings in time order is worse than using findings from the event extraction model in AUC (0.837 event extraction vs. 0.825 LLM). Possible reasons associated with this fact are that the more lengthy LLM extraction contains more findings irrelevant to lung, and that some relevant findings are truncated.

We find that using group-order for LLM-extracted findings decreases performance compared to the time order. One potential reason is that the group-order sequence may lose

Events	AUC	AP	# Input tokens		
			Median	95%	Max
No findings	0.796	0.305	17	27	35
1m	0.833	0.372	38	162	1379
6m	0.838	0.374	42	241	3318
1y	0.825	0.370	47	304	3860
2y	0.830	0.375	53	377	3919
5y (default)	0.837	0.365	62	494	4659

Table 6.7: Effect of different time window lengths for event extraction input sequence, using DistilBERT for risk prediction. We also report the input sequence lengths (in tokens), including the median, 95th percentile, and maximum.

many recent findings during truncation, which are most relevant to the risk.

6.4.5 Impact of Longitudinal Length

We compare the event sequence extracted by the event extraction method using different time windows in Table 6.7, together with demographic information and categorical risk factors including never-smoker, current-smoker, has-COPD, has-emphysema, and has-lung-abnormality. The results show longer inputs do not appear to benefit the risk prediction task beyond the most recent six months. One potential explanation is that older findings may be less relevant to the lung cancer risk and introduce additional noise. One potential reason is that the group-ordered LLM sequence may lose the most recent findings during truncation, which are more likely impactful for risk prediction. We observe that the median length of the group-order sequence exceeds DistilBERT's maximum input length, suggesting severe truncation and the need for another longer context model for risk prediction. Another explanation is that BERT does not model longitudinal dependencies effectively. Because most event sequences fall within the 512-token input limit, input truncation is less likely to be the major reason that affects performance.

	Event sequence (5y)		Event sequence (1m)	
	AUC	AP	AUC	AP
present lung findings (default)	0.837	0.365	0.833	0.372
wo/ anatomy filter	0.820	0.359	0.819	0.357
wo/ assertion filter	0.829	0.361	0.830	0.370

Table 6.8: Effect of including other findings less related to lung cancer risks, from other anatomy or other assertion types, using DistilBERT for risk prediction.

6.4.6 Impact of Less Related Information to Lung Cancer Risk

In Table 6.8, we compare performance when using event sequence with and without filtering to include only present lung findings. The results show that including present findings from other anatomy locations hurts risk prediction. Those other anatomy details are less relevant to lung cancer risk, and may introduce noise. From domain adaptation work lessons, the risk prediction model may learn to use information that is less related to the lung cancer outcome. Even learning on a large dataset, the model can still be distracted. The unfiltered inputs are also longer, so that the model has the risk of truncating critical information. However, the 1 month data, which is shorter, still benefits from filtering irrelevant anatomy.

6.5 Conclusion

We present a lung cancer risk prediction task using risk factors from EHRs. We leverage unstructured EHRs with IE to obtain more complete risk factors, and explore usage of longitudinal information. We create the case-control lung cancer cohort to support this work, as well as other related studies [124]. We find that unstructured data can improve the data completeness of categorical risk factors over a structured EHR source alone, including smoking, COPD, abnormal findings, with corresponding improvements of risk prediction when using a feature-based predictor (XGBoost). A text-based predictor (DistilBERT) obtains similar results to XGBoost using a verbalized representation of the same features,

but it can also incorporate longitudinal lung findings, which further benefits risk prediction performance. The best DistilBERT model, which leverages unstructured data through IE methods, achieves an AUC of 0.84, compared to 0.78 for the baseline model using structured data only.

We find that adding older lung findings is not better than only using a 6-month history when using relatively short event extraction results. There could be multiple reasons for this, it may be that older lung findings are not important to lung cancer risk, or we need a language model to comprehend long contexts, or we simply need a better prompt including the longitudinal representation and instructions. We also did not observe the Llama 3.1 model is better than a much smaller event extraction model for longitudinal lung finding extraction. One potential reason is Llama 3.1 is more verbose and at higher risk of input truncation, as seen in Table 6.6.

We find DistilBERT is sensitive to the representation of the risk factor inputs, adding more clinical information less related to lung cancer risk and changing the chronological order of the lung findings are both harmful. We suspect negative impact from input truncation and therefore plan to evaluate models with longer context windows.

Chapter 7

CONCLUSIONS AND FUTURE WORK

This thesis develops tools that extract information from clinical unstructured data, and applies extracted information for lung cancer risk prediction as a secondary application.

For information extraction tasks, we develop methods for event extraction of symptoms and radiological findings, and longitudinal summarization of radiological findings. Considering clinical data variation and expensive expert annotations, for event extraction, we focus on improving generalization of supervise learning methods under realistic clinical domain shifts. We propose different domain generalization methods for multi-fold domain discrepancy reasons, which can be ensembled for greater benefits. Specifically, we find irrelevant information in the training data hurt domain shifts, and models generalize better after removing unnecessary information to the task within the training data, including masking high-frequency phrases for the encoder-only models, and training with decomposed shorter data for the generative models.

Given that longitudinal EHRs are lengthy, we generate structured summaries using LLMs. We find LLMs can include most critical findings, and GPT-4o can achieve human-level performance in grouping related findings from different reports. However, inclusion of irrelevant information is notable for in Llama 3.1, suggesting potential risk for secondary applications. Future work that reducing verbosity in LLMs using techniques such as reinforcement learning from human feedback (RLHF) methods might help.

In the lung cancer risk prediction task, we find extracted risk factors from clinical reports improve the prediction results over using risk factors solely from the structured EHRs. Extracted radiological findings are effective predictors. When representing longitudinal lung radiological findings as textual inputs, the LLM does not win over much lighter sentence-level IE models, suggesting the advantage of small models when labeled data is available for training. We find irrelevant information about radiological findings not related to lung hurts

prediction, even when training data is substantial, over 30000 examples. We did not observe benefits from using a longer history window beyond 6 months for longitudinal radiological findings, which may explain why the order of the findings matters.

Appendix A

APPENDIX FOR RADIOLOGY FINDING EXTRACTION

A.1 Hierarchical Anatomy Normalization Categories

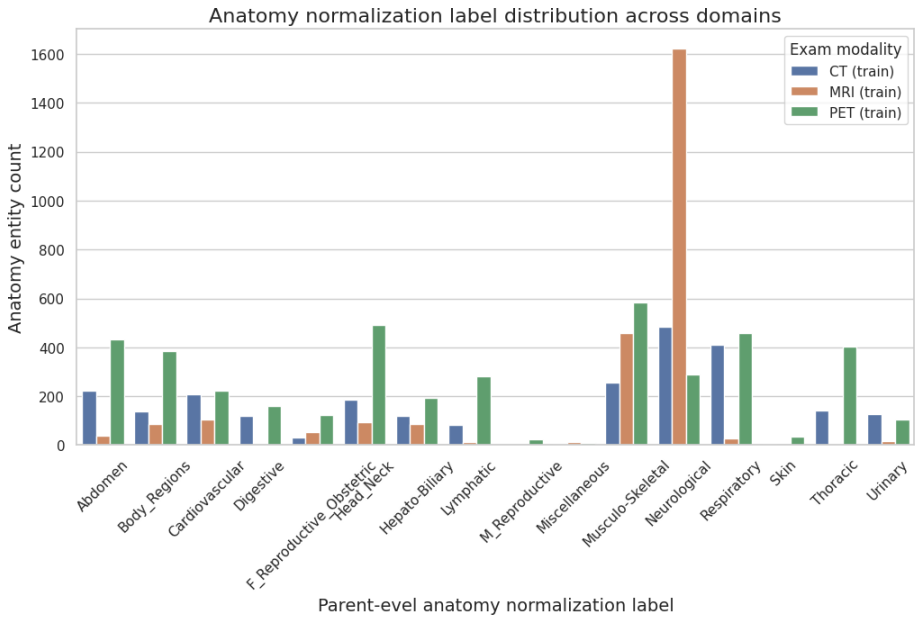

Figure A.1: Domain differences in terms of the frequencies of parent-level anatomy normalization labels from the training data.

We normalize detected anatomy spans for applications focusing on anatomy categories rather than specific anatomy terms. We classify at different granularities, a parent-level coarse classification with 16 parent labels and a child-level fine-grained classification with 72 categories. Each parent-level class includes an "Undetermined" child-level class to account for cases that don't fit into its other specified child classes. The full normalization categories are in Tabel A.1.

As shown in Figure A.1, MRI exhibits a more imbalanced distribution, with a majority

of the anatomies related to the "Neurological" parent-level class. In CT exams, "Respiratory" account for 16% and "Neurological" represent 19% among all finding-related anatomies. For MRI, "Musculo-Skeletal" constitutes 18% while "Neurological" exams make up a substantial 62%. Lastly, in PET, "Head Neck" accounts for 12% and "Musculo-Skeletal" comprises 14%.

A.2 Generative Method Input and Output Formats

We document the templates for the input and output, with examples in Tabel A.2. For the template with contexts, "prepended contexts" include prior sentences, section headers, exam type metadata, other contexts are "appended contexts".

A.3 Post-processing for the Generative Event Extraction

When matching spans in the input sentence for predicted terms, for single-token terms, we match the corresponding token. For multiple-token phrases, we match phrases using the longest common normalized string to the input sentence. Where multiple matches are found, we choose the first match from the left of the sentence, while for anatomy spans, we choose the closest match to their query triggers.

A.4 Domain-level Context Retrieval

We conduct a domain-level context search using 50,000 sentences from the target domain (PET) corpus with more than three tokens, plus 1841 sentences from the test set. The retrieved content must not be the input sentence itself. For each input clinical sentence, we identify the most lexically similar sentence from the search pool by selecting the one with the highest BM25 score. We remove punctuation and lowercase each input query when matching it with the search corpus sentences using the BM25 method.

To filter for anatomy-informative sentences, we employ the same BM25 model to match the entire search corpus with a single anatomy string, which was cheaply curated from the anatomy normalization categories and frequently auto-extracted section headers, as shown in Table A.3. After filtering, the search corpus is reduced to 36%, shrinking from 51,481 sentences to 18,959 sentences.

A.5 Implementation Details

The mSpERT models are trained at a batch size of 15 for 15 epochs.[1] T5 models utilize a maximum input length of 768 tokens and a maximum output length of 512 tokens. When incorporating all types of contexts, we double the input maximum length to 1536 tokens. We train 15 epochs, with a batch size of 8. For the T5 large model, to accommodate a single NVIDIA A100 device, we employ gradient accumulation by using a batch size of 2 and accumulating four times.

A.6 Case study for Context Benefits

We observe that contexts can aid in disambiguation (e.g. right middle lob) and understanding difficult medical terminology (e.g. biapical). For both examples presented in Table A.6, contexts include the term "pulmonary", indicating the anatomies are related to lungs.

[1]We use full event schema for mSpERT models, including all attribute types in the annotations, including anatomy, characteristic, size, size-trend, and count. While T5 models only extract the most important attribute, the anatomy attribute.

Parent-level Class	Child-level Classes
Neurological	Undetermined, Spine Cervical, Spine Thoracic, Spine Lumbar, Spine Sacral, Spine Cord, Spine Unspecified, Brain, Nerve, Pituitary, Cerebrospinal Fluid Pathway, Cerebrovascular System, Extraaxial
Cardiovascular	Undetermined, Venous, Arterial, Pulmonary Artery, Heart, Pericardial Sac, Coronary Artery
Thoracic	Undetermined, Mediastinal
Respiratory	Undetermined, Lung, Pleural Membrane, Tracheobronchial
Digestive	Undetermined, Esophagus, Stomach, Intestine, Small Intestine, Large Intestine
Hepato-Biliary	Undetermined, Gallblader, Bile Duct, Pancreas, Liver
Urinary	Undetermined, Kidney, Urinary Bladder, Ureter
Lymphatic	Undetermined
F Reproductive Obstetric	Undetermined, Breast, Ovary, Uterus, Adnexal, Extra-embryonic, Placenta, Fetus, Umbilical Cord, Female Genital Structure
M Reproductive	Undetermined, Prostate, Testis, Epididymis
Musculo-Skeletal	Undetermined, Skeletal and or Smooth Muscle, Bone and or Joint
Body Regions	Undetermined, Entire Body, Pelvis, Lower Limb, Upper Limb
Head Neck	Undetermined, Thyroid, Neck, Ear, Eye, Mouth, Nasal Sinus, Pharynx, Laryngeal
Skin	Undetermined, Skin and or Mucous Membrane, Subcutaneous
Abdomen	Undetermined, Retroperitoneal, Abdominal Wall, Peritoneal Sac, Spleen, Adrenal Gland, Mesentery
Miscellaneous	Undetermined, Adipose Tissue, Connective Tissue, Biomedical Device

Table A.1: Hierarchical anatomy normalization categories at parent and child levels.

TEMPLATE
Template: [Input sentence] [Question] structured knowledge: [Task ontology] **Template with contexts**: [Prepended contexts] [Input sentence] [Question] structured knowledge: [Appended contexts] [Task ontology]
Trigger task ontology (for T5-vanilla pipeline: 1st step) Indication | Lesion | Medical_Problem **Anatomy task ontology** (for T5-vanilla pipeline: 2nd step, 3rd step) Neurological: Undetermined, Spine Cervical, Spine Thoracic .... (see Table A.1) **Trigger anatomy task ontology** (for all related to one-step building block) trigger types: Indication | Lesion | Medical_Problem anatomy categories: Neurological: Undetermined, Spine, ...
EXAMPLE
Input sentence: 18 x 17 mm hypermetabolic soft tissue density insinuating between the left lobe of the liver and anterior abdominal wall (the R/112) with maximum SUV 14.4 .
Model: T5-vanilla pipeline: first step (trigger span & type) **Question**: Question: What are medical findings in this sentence? **Output**: trigger: density [Lesion]
Model: second step (anatomy span & type) **Question**: Consider the medical finding "density" in the span "hypermetabolic soft tissue density insinuating between the", Question: What anatomy it occurs in? Where is it located? **Output**: anatomies: soft tissue [Hepato-Biliary | Liver], left lobe of the liver [Hepato-Biliary | Liver], anterior abdominal wall [Abdomen | Abdominal Wall]
Model: third step (anatomy normalization) **Question**: Consider the anatomy "soft tissue" in the span "17 mm hypermetabolic soft tissue density insinuating between", which anatomy category it belongs to among listed options? **Output**: anatomies: soft tissue [Hepato-Biliary | Liver]
Model: T5-vanilla one-step (trigger span & type, anatomy span & normalization) **Question**: Question: What are medical findings in this sentence? What anatomy they occur in? which anatomy category they belong to among listed options? **Output**: trigger: density [Lesion] anatomies: soft tissue [Hepato-Biliary | Liver], left lobe of the liver [Hepato-Biliary | Liver], anterior abdominal wall [Abdomen | Abdominal Wall]
Model: T5 one-step subtask blocks (trigger span & type, anatomy span & normalization) **Question**: [same as T5-vanilla one-step] **Output**: state: trigger detection answer: density state: trigger classification answer: density [Lesion] state: span detection answer: soft tissue, left lobe of the liver, anterior abdominal wall state: classification answer: soft tissue [Hepato-Biliary | Liver] state: classification answer: left lobe of the liver [Hepato-Biliary | Liver] state: classification answer: anterior abdominal wall [Abdomen | Abdominal Wall]
Model: multitask for trigger classification (trigger type) **Question**: Consider the medical finding "density", Question: What is the type of this medical finding? **Output**: state: trigger classification answer: density [Lesion]
Model: multitask for anatomy span (anatomy span) **Question**: Consider the medical finding "density" in the span "hypermetabolic soft tissue density insinuating between the", Question: Please identify terms that describe the finding's anatomy locations. **Output**: state: span detection answer: soft tissue, left lobe of the liver, anterior abdominal wall

Table A.2: Templates and examples for T5 inputs and outputs. The "multitask" rows correspond to auxiliary tasks for the T5 one-step subtask block method. We omit rows for "multitask for anatomy" and "multitask for anatomy normalization", since they use the same question format as the 2nd and 3rd steps of the pipeline approach, but with answers in the subtask block format.

Neurological: Spine Cervical, Spine Thoracic, Spine Lumbar, Spine Sacral, Spine Cord, Spine, Brain, Nerve, Pituitary, Cerebrospinal, Cerebrovascular, Extraaxial
Cardiovascular: Venous, Arterial, Pulmonary Artery, Heart, Pericardial Sac, Coronary Artery
Thoracic: Mediastinal
Respiratory: Lung, Pleural Membrane, Tracheobronchial
Digestive: Esophagus, Stomach, Intestine, Intestine, Intestine
Hepato-Biliary: Gallbladder, Bile, Pancreas, Liver
Urinary: Kidney, Urinary Bladder, Ureter
Reproductive: Breast, Ovary, Uterus, Adnexal, Extra-embryonic, Placenta, Fetus, Umbilical Cord, Genital Structure, Prostate, Testis, Epididymis
Musculo-Skeletal: Skeletal, Smooth Muscle, Bone, Pelvis, Limb
Head Neck: Thyroid, Neck, Ear, Eye, Mouth, Nasal Sinus, Pharynx, Laryngeal
Skin: Skin, Mucous Membrane, Subcutaneous
Abdomen: Retroperitoneal, Abdominal, Peritoneal Sac, Spleen, Adrenal, Mesentery, Adipose, Chest, Mediastinum, Osseous, Bones, Extremities, Lungs, Musculoskeletal, Ventricular, Bowel, Pleura, Spleen, Vasculature, Thorax, Gallbladder, Kidneys, Adrenals, Adrenal, Cardio

Table A.3: Common anatomy terms for filtering the search scope of domain-level context retrieval. This list is curated from the anatomy task ontology (Table A.1) and frequent section headers. Stop words are removed.

Error with example	Contexts	Before and after
[ambiguity] **Right middle lobe** nodule (4, 81) measures 3 mm, previously 4 mm	[document-level section header] Scattered bilateral **pulmonary** nodules, as described below	before: Hepato-Biliary | Liver after: Respiratory | Lung
[hard vocabulary] There is **biapical** fibrosis	[domain-level BM25] There is biapical **pulmonary** fibrosis compatible with radiation therapy	before: Musculo-Skeletal | Bone and or Joint after: Respiratory Lung

Table A.4: Error examples with helpful contexts

Appendix B

APPENDIX FOR TIMELINE GENERATION

B.1 Annotation Guidelines

A	B	C	D	E	F
Rowid / group ID	Finding_group	[DATE1] _chest_ct	group ID	[DATE2] _chest_ct	group ID
1	Right upper lobe ground glass nodule	**10.2 mm ground glass nodule in the posterior segment of the right upper lobe with a small central nodular component**	1	**10 mm ground glass nodule in the right upper lobe with a central 4 mm cavity**	1
2	Right lower lobe ground glass nodule	2.0 cm ground glass opacity in the posterior segment of the right lower lobe with a small central nodular component	**2**	2 cm ground glass opacity in the posterior basal segment of the right lower lobe	2
3	Right upper lobe subsolid nodule	**2.6 mm noncalcified nodule in the anterior segment of the left upper lobe**	5	**4 mm subsolid nodule in the right upper lobe**	Skip if rated as wrong
4	Right lower lobe subsolid nodule				
5	Sub-3 mm nodule in the anterior upper lingula	Sub-3 mm nodule in the anterior upper lingula	Skip if rated as wrong	Sub-3 mm nodule in the anterior upper lingula	**5**
6	**Cardiomegaly**	**cardiomegaly**	Skip if rated as irrelevant		
7	**(new group name if need)**				

Figure B.1: A group assignment example provided to annotators.

The annotation pipeline consists of three sequential tasks. In the first task (Table B.1), each cell is rated for factuality into levels of -1, 0, 1, 2. The second task (Table B.2), collects final task group assignment on cells to form the timeline rows. Figure B.1 shows one example presented to annotators. Finally, the third task (Table B.3) asks an annotator to revise the group names on formed groups.

B.2 Prompts

The prompts used in the three-step pipeline are provided in the subsequent tables. The step 1 prompt for finding extraction is shown in Table B.4, and the step 2 group name generation prompt is presented in Table B.5.

Step 3 group assignment prompts have three versions: single finding (Table B.6), multi-finding (Table B.8), and multi-finding without group names provided (Table B.7).

B.3 Qualitative Analysis

We present LLM generation examples for Step 1 finding extraction in Table B.9 and Step 2 group name generation in Table B.10 and Table B.11. In both steps, GPT-4o generates fewer findings with greater conciseness, while Llama 3.1 tends to generate more findings and include content irrelevant to lung radiological findings. In Step 2, when switching from full longitudinal report context to the predicted finding list as a shorter context, GPT-4o produces more group names, possibly because the full context helps merge findings. However, this trend of reduced group names is not observed in Llama 3.1, potentially because the shorter context instead reduces irrelevant content in the Llama output.

Task 1: Cell-level Factuality Rating

- For all cells in the column, rate the factuality level [0 bad, 1 partially correct, 2 good] for each cell based on the consistency with the column-associated report.
- Annotators should focus only on finding-related facts written in the specific report being evaluated. Do not refer to facts from other reports.
- Correctness of those facts should **NOT** affect the rating:
  - Suggestive conditions
    - e.g., "considering neoplasm given the persistent nodules"
    - e.g., "likely due to infection"
  - Facts that are not lung-related or not radiological findings
    - e.g., "Lung RAD rating is 3A"
    - e.g., "Coronary Artery Calcifications"
- **Skip 'not-mentioned' cells.** Blank cells are equivalent to 'not mentioned'. Instead of annotating them, add them in Task 2 for missing findings if actually mentioned.
- Label the cell as not containing lung-related findings as [**-1: non-lung-findings**].
- Add a note column next to the factuality cell.

Table B.1: Annotation guidelines for factuality rating

Task 2: Group Assignments

Assign a **group label** for each cell. The group labels to choose are the headers in the `finding_group` column. You may **add a new group label** if no existing group labels can distinguish this cell from other groups.

What should be included in a group

- The same finding in the notes, even if described differently, **or**
- Findings from different notes that can change to one another.

This may lead to trend descriptions such as:

- new
- increasing
- resolving

Duplicate labels If two group labels are duplicated, **choose one label and use it consistently** for all cells belonging to that group.

Goal:

Ensure that the same findings or time-related findings are placed in the same group.

Skip condition:

Skip cells where factuality is 0 or -1.

Table B.2: Annotation guidelines for group assignments

Task 3: Group Name Revision

Given a group of dated findings, add the group name for this group. The goal is to make the group name clearly distinguishable from other groups while remaining concise and containing critical information that a radiologist might use when referring to the finding.

Strategy:

The group name should:

1. **Include the facts shared by all members in the group.** When paraphrasing occurs, choose the description used more frequently.
2. **Omit details that can change in the future** (e.g., size, change trend, severity).
3. Use extra details if needed to distinguish similar groups.

Table B.3: Annotation guidelines for group name revision given formed groups.

Prompt for step 1 finding extraction

Q: [**Single_Radiology_Report_Text**, e.g. EXAMINATION: CT of the chest without contrast - low dose lung cancer screening.....]:

Please list all lung-related findings (including pleura-ralated) from the exam report. Please use a python list to include the strings of those lung desription. Each string should include all details related to higher or lower risk, using the same description in the note , especially finding's location, size, trend of change, and shape, density, or other characteristics.

Good example: ["stable 2 mm left upper lobe calcified nodule", 'left lower lobe opacities']

Now please anwswer:

lung_finding_list =

A: Here is the list of lung-related findings from the exam report:

"'python lung_finding_list = ["2.6 mm noncalcified nodule in the anterior segment of the left upper lobe", "alveolar groundglass opacity in the posterior segment of the right upper lobe with a small central nodular component measuring 10.2 mm", ...

Table B.4: Step 1 finding extraction input. The prompt includes the single report input, an one-shot example, and instructions requiring a python list output format.

Prompt for step 2 group name generation
Q: [**Longitudinal_Textual_Reports**, e.g. date,note_tag,QOClassificationName,ReportText YYYY-MM-DD,chest_ct,CT Lung Low Dose Screening,"EXAMINATION: CT of the chest without contrast - low dose lung cancer screening ...]:
Please group the lung-related findings recorded over time (including pleura-related findings) and list the group names.
Each group should contain findings that are essentially the same, even if described differently, as well as findings that evolve into one another over time.
Different findings or findings will not evolve into one another should be in separate groups
Group names should be specific that separate unrelated findings, avoid generic names.
Good group name: right upper lobe ground glass nodule
bad group name: ground glass nodules
Finally give me a list of group names separated by commas ’,’
A: .... List of group names: Right upper lobe ground glass nodule, Right lower lobe ground glass opacity with central nodular component, ...

Table B.5: Step 2 group name generation prompt that requires list group names at the end of response. The prompt includes a good and a bad example for individual group names, but no examples for the list format.

Prompt for step 3 group assignment for single finding

Q: [**Longitudinal_Textual_Reports**, e.g. date,note_tag,QOClassificationName,ReportText YYYY-MM-DD,chest_ct,CT Lung Low Dose Screening,"EXAMINATION: CT of the chest without contrast - low dose lung cancer screening ...]
Please find the group best describing this finding given the radiology exam history and group labels.

There are groups presented to track the lung-related findings recorded over time.
Each group should contain findings that are essentially the same, even if described differently, as well as findings that evolve into one another over time.
Different findings or findings will not evolve into one another should be in separate groups

[**Zero-Shot Content**]
The group labels are:

[**Group name list from step 2:** e.g. Right upper lobe ground glass nodule, Right lower lobe ground glass nodule, ...], other
The finding to be classified is: [**Single finding from step 1:** e.g. alveolar groundglass opacity in the posterior segment of the right upper lobe with a small central nodular component measuring 10.2 mm]

Give me the group in the last line starting with 'Answer:'.

[**Few-Shot Content**]
Here are some examples:

Example 1

Group labels: small RUL nodule, 10 mm right lower lobe nodule, other

Finding to be classified: Small 4 mm right upper lobe nodule stable

Answer: small RUL nodule

Example 2

Group labels: calcified granuloma, right upper lobe nodules, other

Finding to be classified: Benign appearing calcified granuloma in left lung

Answer: calcified_granuloma

Your task:

Group labels: [**Group name list from step 2:** e.g. Right upper lobe ground glass nodule, Right lower lobe ground glass nodule, ...], other

Finding to be classified: [**Single finding from step 1:** e.g. alveolar groundglass opacity in the posterior segment of the right upper lobe with a small central nodular component measuring 10.2 mm]

Your explain: [EXPLAIN WHY]

Answer: [GROUP_NAME CHOICE]

Always end with "Answer: [GROUP_NAME]" as the very last line
A: ... Answer: Right upper lobe ground glass nodule

Table B.6: Step 3 Group Assignment for Single Finding. The prompts include underlined texts only when using full-contexts. We have both zero-shot and few-shot instructions, few-shot version adds two examples and further emphasize on the format requirements.

Prompt for step 3 group assignment (multi-finding tag format, no group names given)
Q: Task: Your task is to group radiology findings evolving over time from multiple reports of the same patient. Please replace every <#> with a group name tag in angle-bracket format, such as <10_mm_right_lower_lobe_nodule>. You should invent your own group tags so that each tag is representative for its associated group. Use the same group name tag across all reports for all mentions related to the same underlying entity. Your Task: Radiology Findings: [**Finding_List_with_Markers**, e.g. 2.6 mm noncalcified nodule in the anterior segment of the left upper lobe <#> alveolar groundglass opacity in the posterior segment of the right upper lobe with a small central nodular component measuring 10.2 mm <#> ...] Radiology Findings END Please apply a group name tag consistently to all members of the group. Please reply with exactly the same lines listed above that are between 'Radiology Findings' and 'Radiology Findings END', but with the group name tags inserted.

Table B.7: Step 3 Group Assignment for Multi-Finding when no group names are given. The prompt includes an example of single tag, but no examples for the final output for all provided findings.

Prompt for step 3 group assignment for multi finding

Q: [**Longitudinal_Textual_Reports**, e.g. date,note_tag,QOClassificationName,ReportText
YYYY-MM-DD,chest_ct,CT Lung Low Dose Screening,"EXAMINATION: CT of the chest without contrast - low dose lung cancer screening ...]
Task: Your task is to group radiology findings evolving over time from multiple reports of the same patient. Those reports are as given above.
Please replace every <#> with a group name tag in angle-bracket format, such as <10_mm_right_lower_lobe_nodule>.
You can use either an existing group tag as provided, or invent your own.
Use the same group name tag across all reports for all mentions related to the same underlying entity.
[**Few-Shot Content**]
Here are examples:
Example 1
Existing group name tags:
<10_mm_right_lower_lobe_nodule> : 10 mm right lower lobe nodule
<small_RUL_nodule> : small RUL nodule
Radiology Findings:
Small 4 mm right upper lobe nodule is new <#>
The 10 mm right lower lobe groundglass is stable <#>
Here are several left lobe calcified nodules <#>
Small right upper lobe nodule is stable <#>
RLL groundglass is smaller now <#>
Radiology Findings END
Answer:
Small 4 mm right upper lobe nodule is new <small_RUL_nodule>
The 10 mm right lower lobe groundglass is stable <10_mm_right_lower_lobe_nodule>
Here are several left lobe calcified nodules <left_lobe_calcified_nodules>
Small right upper lobe nodule is stable <small_RUL_nodule>
RLL groundglass is smaller now <10_mm_right_lower_lobe_nodule>
Example 2
Existing group name tags:
<calcified_granuloma> : calcified granuloma
<right_upper_lobe_nodules> : right upper lobe nodules
Radiology Findings:
Benign appearing calcified granuloma in the left lung <#>
New right upper lobe nodules compared to last scan <#>
A 1.5 cm part-solid nodule is identified in the lingula <#>
Radiology Findings END
Answer:
Benign appearing calcified granuloma in the left lung <calcified_granuloma>
New right upper lobe nodules compared to last scan <right_upper_lobe_nodules>
A 1.5 cm part-solid nodule is identified in the lingula <part_solid_nodule_in_the_lingula>
[**Few-Shot Content END**]
Your Task:
Existing group name tags:
[**Group name list from step 2:** e.g. right_upper_lobe_ground_glass_nodule, right_lower_lobe_ground_glass_nodule, ...]
Radiology Findings:
[**Finding_List_with_Tags**,
e.g. 2.6 mm noncalcified nodule in the anterior segment of the left upper lobe <#>
alveolar groundglass opacity in the posterior segment of the right upper lobe with a small central nodular component measuring 10.2 mm <#> ...]
Radiology Findings END
If an existing group name tag fits, use it;
If multiple existing tags fit, choose one and apply it consistently to all members of the group.
If no existing group name fits, invent your own group tag.
Please reply with exactly the same lines listed above that are between 'Radiology Findings' and 'Radiology Findings END', but with the group name tags inserted.
Answer:

Table B.8: Step 3 Group Assignment for Multi-Finding when group names are given. The prompts include underlined texts only when using full-contexts. We have both zero-shot and few-shot instructions, few-shot version adds two examples.

Gold	GPT-4o	Llama3.1
Moderate centrilobular emphysema stable lung nodules since [DATE], therefore benign. No new or enlarging pulmonary nodule	Moderate centrilobular emphysema Stable lung nodules since [DATE], therefore benign No new or enlarging pulmonary nodule	Moderate centrilobular emphysema stable lung nodules since [DATE], therefore benign. No new or enlarging pulmonary nodule Normal supraclavicular region Normal mediastinum Normal pleura Mildly dilated main pulmonary artery measuring 3.3 cm in caliber, also unchanged Stable mildly dilated pulmonary artery may represent pulmonary hypertension Stable mildly dilated ascending aorta may represent systemic hypertension
Lungs: A few tiny (1–2 mm) nodules are scattered throughout the lungs [IMAGE NUMBES].	No large masses or dense airspace consolidations A few tiny (1–2 mm) nodules are scattered throughout the lungs	Supraclavicular region: Normal. Mediastinum: Normal. Lungs: No large masses or dense airspace consolidations. Lungs: A few tiny (1–2 mm) nodules are scattered throughout the lungs. Pleura: Normal.
Low lung volume Left base atelectasis	low lung volume with left base atelectasis no evidence of edema no pneumothorax	Cardiomegaly Low lung volume Left base atelectasis No evidence of edema No pneumothorax
Right basilar consolidation likely pneumonia Patchy consolidation at the left lung base Left basilar consolidation likely pneumonia Minimal density at right lung base	patchy consolidation at the left lung base minimal density at right lung base	Patchy consolidation at the left lung base Minimal density at right lung base Right basilar consolidation likely pneumonia Left basilar consolidation likely pneumonia
predominant fibrosis as before increased opacification of the left lower lobe	predominant fibrosis increased opacification of the left lower lobe best seen in the lateral view which may represent atelectasis, aspiration, or pneumonia	predominant fibrosis increased opacification of the left lower lobe atelectasis, aspiration, or pneumonia (left lower lobe) no pleural effusion no pneumothorax

Table B.9: Step 1 finding extraction examples from GPT-4o and Llama 3.1 8B Instruct. Llama 3.1 tends to generate more findings, and to include content irrelevant to lung findings. Human references are provided as the gold labels. The predicted findings are presented in the same order as they appear in the LLM responses.

Gold (group assignments from Ann1)	**Gold (group assignments from Ann2)**	**GPT-4o**	**Llama3.1**
Hyperinflated lungs	Centrilobular emphysema in upper lobe predominant distribution	Left upper lobe spiculated nodule	Centrilobular emphysema in upper lobe predominant distribution
Flattening of domes of diaphragm	Focal area of scarring	Centrilobular emphysema	Spiculated nodule in left upper lobe (concerning for primary lung malignancy)
Centrilobular emphysema in upper lobe predominant distribution	Spiculated nodule in left upper lobe (concerning for primary lung malignancy)	Pleura and pleural spaces	Unchanged spiculated left upper lobe nodule (6 x 9 mm)
Left anterior lingula scarring	Right upper lobe nodule	Pericardial effusion	Unchanged right upper lobe nodule (2 mm)
Spiculated nodule in left upper lobe (concerning for primary lung malignancy)		Ascending aorta abnormalities	No new pulmonary nodules
Right upper lobe nodule			Pleural effusions or pneumothorax (no findings)
Bilateral upper lobe lung nodules			Small bilateral fat-containing Bochdalek hernias
			Mild enlargement of the left lobe of the thyroid (supraclavicular region)
			Normal heart size
			Trace pericardial effusion (likely physiologic)
			Ascending aorta ectasia (unchanged)
			Mildly dilated main pulmonary artery (stable)
			Nonobstructive bilateral punctate renal calculi (upper abdomen)
			Moderate centrilobular emphysema
			Large lung volumes consistent with emphysema
			No acute bone or soft tissue abnormality

Table B.10: Step 2 group name generation examples from GPT-4o and Llama 3.1 8B Instruct when using full longitudinal reports as the context. Gold group names are provided in two versions, each reflecting the grouping assignments of Ann1 and Ann2 respectively, with names refined by a single third annotator across both versions. Similarly to step 1 results, Llama 3.1 tends to generate more group names and to include content irrelevant to lung findings. Two versions of human references are provided as the gold labels. The predicted group names are presented in the same order as they appear in the LLM responses.

GPT-4o (Full)	GPT-4o (List)	Llama3.1 (Full)	Llama3.1 (List)
Left upper lobe spiculated nodule	Hyperinflated lungs	Centrilobular emphysema in upper lobe predominant distribution	Hyperinflated lungs and diaphragm flattening
Centrilobular emphysema	Flattening of the domes of the diaphragm	Spiculated nodule in left upper lobe (concerning for primary lung malignancy)	Centrilobular emphysema
Pleura and pleural spaces	Centrilobular emphysema	Unchanged spiculated left upper lobe nodule (6 x 9 mm)	Spiculated nodule in left upper lobe (primary lung malignancy concern)
Pericardial effusion	Spiculated nodule in the left upper lobe	Unchanged right upper lobe nodule (2 mm)	Small focal area of scarring in lingula
Ascending aorta abnormalities	Right upper lobe pulmonary nodule	No new pulmonary nodules	Clear trachea and central airways
	Small focal area of scarring within the anterior aspect of the lingula	Pleural effusions or pneumothorax (no findings)	Pleural effusions or pneumothorax (absence of)
	No pleural effusions or pneumothorax	Small bilateral fat-containing Bochdalek hernias	Bilateral fat-containing Bochdalek hernias
	Small bilateral fat-containing Bochdalek hernias	Mild enlargement of the left lobe of the thyroid (supraclavicular region)	Large lung volumes consistent with emphysema
	Large lung volumes consistent with emphysema	Normal heart size	Stable lung nodules (benign)
	The lungs and pleural spaces are clear	Trace pericardial effusion (likely physiologic)	Normal supraclavicular region and mediastinum
		Ascending aorta ectasia (unchanged)	Normal pleura
		Mildly dilated main pulmonary artery (stable)	Mildly dilated main pulmonary artery (pulmonary hypertension concern)
		Nonobstructive bilateral punctate renal calculi (upper abdomen)	Mildly dilated ascending aorta (systemic hypertension concern)
		Moderate centrilobular emphysema	
		Large lung volumes consistent with emphysema	
		No acute bone or soft tissue abnormality	

Table B.11: Comparison of step 2 group name generation examples using full longitudinal reports versus predicted finding lists as context. Examples are from GPT-4o and Llama 3.1 8B Instruct. The predicted group names are presented in the same order as they appear in the LLM responses.

Appendix C

APPENDIX FOR LUNG CANCER PREDICTION

C.1 Prompts for LLM Risk Factor Extraction

User The patient also havs clincal notes describing smoking use as below: [clinical note content] Question: Consider both the structured smoking records and the clinical notes sentences, Please answer is this patient a current smoker, a quit smoker who smoke in the past, or a never smoker ? please first explain with evidence and finally answer with one of the choice [current|quit|never|unknown]

Table C.1: Prompts of smoking status extraction from clinical notes

The patient and has clincal notes as below orderd from earlier to the latest:
[related-sentences]
Question:
Consider the clinical notes sentences, I am looking for Chronic Pulmonary Disease COPD disease history.
COPD means Chronic obstructive pulmonary disease (COPD). COPD often include emphysema and chronic bronchitis.
COPD is defined as lung condition characterized
by chronic respiratory symptoms (dyspnea, cough, sputum production and/or exacerbations) due to abnormalities of the airways (bronchitis, bronchiolitis) and/or alveoli (emphysema) that cause persistent, often progressive, airflow obstruction.
please answer the following questions:
If the patient has any COPD (including emphysema, chronic bronchitis, or any COPDs)? (answer with has_copd = yes or has_copd = no).
If the patient has any emphysema? (answer with has_emphysema = yes).
If the patient has any chronic bronchitis? (answer with has_chronic_bronchitis = yes or has_chronic_bronchitis = no).
First, answer with the size that is explicitly mentioned,
for example,
for note sentencs " Mild centrilobular emphysema is identified and stable",
answer:
has_copd = yes
has_emphysema = yes
has_chronic_bronchitis = no
for " this patient has COPDs",
answer:
has_copd = yes
has_emphysema = no
has_chronic_bronchitis = no
Please first select evidence sentences related to COPDs, select no more than 6 sentences.
then answer with "has_copd= [yes|no] has_emphysema= [yes|no] has_chronic_bronchitis= [yes|no] "for each evidence"
Finally summarize above answers into one responsse
Finally answer with the format in the end of response using with a single binary yes|no (the best overall answer) per slot. Please answer with yes or no
Final Answer: has_copd= , has_emphysema= , has_chronic_bronchitis= ,

Table C.2: Prompts of COPD extraction

The patient and has clincal notes as below orderd from earlier to the latest:
[related-sentences]
The current date is [current-date].
Question:
Consider the clinical notes sentences, I am looking for lung nodules related to lung cancer risk. please answer the following questions:
If the patient has any pulmonary nodule, what is the most recent pulmonary size in mm? (answer with any_pulmonary_nodule_size= xx mm).
If the patient has a solid lung nodule (also as solitary nodule, etc), what is the most recent solid nodule size in mm? (answer with solid_nodule_size= xx mm).
If the patient has a part-solid lung nodule (mentioned as subsolid, etc), what is the most recent part solid nodule size in mm? (answer with part_solid_nodule_size= xx mm).
If the patient has a non-solid nodule (also mentioned as GGN, Ground-glass opacity nodules, etc), what is the most recent non-solid(or GGN) nodule size in mm? (answer with non_solid_or_ggn_nodule_size= xx mm).
First, answer with the size that is explicitly mentioned,
for example, for " a 3 mm pulmonary nodule", any_pulmonary_nodule_size=3 mm
for " a 5*6 mm solid nodule", solid_nodule_size=6 mm
Answer with the largest diameter when multiple dimensions are reported.
Please first select no mor than 6 evidence sentences related to lung nodule size.
then answer with "solid_nodule_size= NUMBER mm for each evidence"
Finally summarize above answers into one responsse
Finally answer with the format in the end of response using with a single number/unknown (the best overall answer) per slot.
Final Answer: any_pulmonary_nodule_size= xx mm, solid_nodule_size= xx mm, part_solid_nodule_size= xx mm non_solid_or_ggn_nodule_size= xx mm,

Table C.3: Prompts of abnormal lung finding nodule size.

Summarize the lung-related findings and how they change over time from the reports below. Please pay attention to lung-related findings.

[radiology reports]

Please summarize lung-related findings from chest CT and X-ray reports, as well as their temporal changes if any. Common findings to consider include: nodule, opacity, consolidation, atelectasis, pleural effusion, masses, fibrosis, pneumonia, low lung volume, volume loss, adenopathy, interstitial changes, infiltrates, tuberculosis, emphysema, and bronchitis. First, please list evidence about abnormal lung findings chronologically from oldest to most recent, labeling with time and exam type. These findings may come from the common categories listed above or may include other findings outside of this list. Please include important finding details if available, such as location, size, shape, and density, etc.

Next, for each finding, describe their temporal changes based on evidence, using terms like "new", "worsen", "persistent", "stable", "improved", "resolved".

Finally, conclude with a concise paragraph summarizing all abnormal lung findings, including important details and any temporal changes where

Table C.4: Prompts of LLM lung finding summarization

BIBLIOGRAPHY

[1] Bin Ji, Jie Yu, Shasha Li, Jun Ma, Qingbo Wu, Yusong Tan, and Huijun Liu. Span-based joint entity and relation extraction with attention-based span-specific and contextual semantic representations. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 88–99, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.8. URL `https://aclanthology.org/2020.coling-main.8/`.

[2] Kevin Lybarger, Aashka Damani, Martin L. Gunn, Özlem Uzuner, and Meliha Yetisgen-Yildiz. Extracting radiological findings with normalized anatomical information using a span-based BERT relation extraction model. In *AMIA Annu Symp Proc.*, 2022.

[3] Kevin Lybarger, Mari Ostendorf, M. Thompson, and Meliha Yetisgen. Extracting COVID-19 diagnoses and symptoms from clinical text: A new annotated corpus and neural event extraction framework. In *J Biomed Inform*, 2021.

[4] Kevin Lybarger, Nicholas J Dobbins, Ritche Long, Angad Singh, Patrick Wedgeworth, Özlem Uzuner, and Meliha Yetisgen. Leveraging natural language processing to augment structured social determinants of health data in the electronic health record. *J Am Med Inform Assoc.*, 2023.

[5] Qiuhao Lu, Dejing Dou, and Thien Nguyen. ClinicalT5: A generative language model for clinical text. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5436–5443, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.findings-emnlp.398`.

[6] Yun Shen, Jiamin Yu, Jian Zhou, and Gang Hu. Twenty-five years of evolution and hurdles in electronic health records and interoperability in medical research: Comprehensive review. *Journal of Medical Internet Research*, 27:e59024, January 2025. ISSN 1438-8871. doi: 10.2196/59024. URL `https://www.jmir.org/2025/1/e59024`.

[7] Maryam Tayefi, Phuong Ngo, Taridzo Chomutare, Hercules Dalianis, Elisa Salvi, Andrius Budrionis, and Fred Godtliebsen. Challenges and opportunities beyond structured data in analysis of electronic health records. *WIREs Computational Statistics*, 13(6):e1549, November 2021. ISSN 1939-5108, 1939-0068. doi: 10.1002/wics.1549. URL `https://wires.onlinelibrary.wiley.com/doi/10.1002/wics.1549`.

[8] Tom M Seinen, Jan A Kors, Erik M Van Mulligen, and Peter R Rijnbeek. Using structured codes and free-text notes to measure information complementarity in electronic health records: Feasibility and validation study. *Journal of Medical Internet Research*, 27:e66910, February 2025. ISSN 1438-8871. doi: 10.2196/66910. URL `https://www.jmir.org/2025/1/e66910`.

[9] Kevin Lybarger, Namu Park, Sitong Zhou, Aashka Damani, Alison Brennan, Jagjeet Gill, Nianiella Dorvall, Vy Huynh, Spencer Lewis, Martin L. Gunn, Özlem Uzuner, and Meliha Yetisgen-Yildiz. A corpus of radiology reports from multiple imaging modalities with fine-grained event-based annotations. In *AMIA Annu Symp Abstract*, 2022.

[10] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large language models are few-shot clinical information extractors. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.130. URL `https://aclanthology.org/2022.emnlp-main.130`.

[11] Yujuan Fu, Giridhar Kaushik Ramachandran, Nicholas J. Dobbins, Namu Park, Michael Leu, Abby R. Rosenberg, Kevin Lybarger, Fei Xia, Özlem Uzuner, and

Meliha Yetisgen. Extracting social determinants of health from pediatric patient notes using large language models: Novel corpus and methods. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7045–7056, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.618/.

[12] Qingyu Chen, Yan Hu, Xueqing Peng, Qianqian Xie, Qiao Jin, Aidan Gilson, Maxwell B. Singer, Xuguang Ai, Po-Ting Lai, Zhizheng Wang, Vipina K. Keloth, Kalpana Raja, Jimin Huang, Huan He, Fongci Lin, Jingcheng Du, Rui Zhang, W. Jim Zheng, Ron A. Adelman, Zhiyong Lu, and Hua Xu. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature Communications*, 16(1): 3280, April 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-56989-2. URL https://www.nature.com/articles/s41467-025-56989-2.

[13] Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, Kirk Roberts, and Hua Xu. Improving large language models for clinical named entity recognition via prompt engineering. *J Am Med Inform Assoc.*, 31(9):1812–1820, September 2024. ISSN 1067-5027, 1527-974X. doi: 10.1093/jamia/ocad259. URL https://academic.oup.com/jamia/article/31/9/1812/7590607.

[14] Yanis Labrak, Mickael Rouvier, and Richard Dufour. A zero-shot and few-shot study of instruction-finetuned large language models applied to clinical and biomedical tasks. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2049–2066, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.185/.

[15] R. Verma, E. Alsentzer, Z. Strasser, L. Chang, K. Roman, E. Gershanik, C. Hernandez, M. Linares, J. Rodriguez, D. Thakral, O. Unlu, J. You, L. Zhou, and D. Bates. Verifiable summarization of electronic health records using large language models to support chart review. *medRxiv*, June 2025. doi: 10.1101/2025.06.02.25328807. URL https://doi.org/10.1101/2025.06.02.25328807. Preprint.

[16] Wei-Qi Wei, Pedro L Teixeira, Huan Mo, Robert M Cronin, Jeremy L Warner, and Joshua C Denny. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J Am Med Inform Assoc.*, 23(e1):e20–e27, April 2016. ISSN 1527-974X, 1067-5027. doi: 10.1093/jamia/ocv130. URL https://academic.oup.com/jamia/article/23/e1/e20/2379841.

[17] Preethi Raghavan, James L. Chen, Eric Fosler-Lussier, and Albert M. Lai. How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? *AMIA Jt Summits Transl Sci Proc*, 2014:218–223, 2014. ISSN 2153-4063.

[18] Peter L Elkin, Sarah Mullin, Jack Mardekian, Christopher Crowner, Sylvester Sakilay, Shyamashree Sinha, Gary Brady, Marcia Wright, Kimberly Nolen, JoAnn Trainer, Ross Koppel, Daniel Schlegel, Sashank Kaushik, Jane Zhao, Buer Song, and Edwin Anand. Using artificial intelligence with natural language processing to combine electronic health record's structured and free text data to identify nonvalvular atrial fibrillation to decrease strokes and death: Evaluation and case-control study. *Journal of Medical Internet Research*, 23(11):e28946, November 2021. ISSN 1438-8871. doi: 10.2196/28946. URL https://www.jmir.org/2021/11/e28946.

[19] Kim Huat Goh, Le Wang, Adrian Yong Kwang Yeow, Hermione Poh, Ke Li, Joannas Jie Lin Yeow, and Gamaliel Yu Heng Tan. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nature Communications*, 12 (1):711, January 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-20910-4. URL https://www.nature.com/articles/s41467-021-20910-4.

[20] Kenneth L. Kehl, Justin Jee, Karl Pichotta, Morgan A. Paul, Pavel

Trukhanov, Christopher Fong, Michele Waters, Ziad Bakouny, Wenxin Xu, Toni K. Choueiri, Chelsea Nichols, Deborah Schrag, and Nikolaus Schultz. Shareable artificial intelligence to extract cancer outcomes from electronic health records for precision oncology research. *Nature Communications*, 15(1):9787, November 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-54071-x. URL https://www.nature.com/articles/s41467-024-54071-x.

[21] John C. Ruckdeschel, Mark Riley, Sriram Parsatharathy, Rajesh Chamarthi, Chakethraman Rajagopal, Hui Shuang Hsu, Doug Mangold, and Chiny Driscoll. Unstructured data are superior to structured data for eliciting quantitative smoking history from the electronic health record. *JCO Clinical Cancer Informatics*, (7): e2200155, February 2023. ISSN 2473-4276. doi: 10.1200/CCI.22.00155. URL https://ascopubs.org/doi/10.1200/CCI.22.00155.

[22] Honghan Wu, Giulia Toti, Katherine I Morley, Zina M Ibrahim, Amos Folarin, Richard Jackson, Ismail Kartoglu, Asha Agrawal, Clive Stringer, Darren Gale, Genevieve Gorrell, Angus Roberts, Matthew Broadbent, Robert Stewart, and Richard Jb Dobson. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. *J Am Med Inform Assoc.*, 25(5):530–537, May 2018. ISSN 1067-5027, 1527-974X. doi: 10.1093/jamia/ocx160. URL https://academic.oup.com/jamia/article/25/5/530/4817428.

[23] Shivani Mehta, Courtney R Lyles, Anna D Rubinsky, Kathryn E Kemper, Judith Auerbach, Urmimala Sarkar, Laura Gottlieb, and William Brown Iii. Social determinants of health documentation in structured and unstructured clinical data of patients with diabetes: Comparative analysis. *JMIR Med Inform.*, 11: e46159–e46159, August 2023. ISSN 2291-9694. doi: 10.2196/46159. URL https://medinform.jmir.org/2023/1/e46159.

[24] Maria Giulia Prado, Larry G. Kessler, Margaret A. Au, Hannah A. Burkhardt, Monica L. Zigman Suchsland, Lesleigh Kowalski, Kari A. Stephens, Meliha

Yetisgen, Fiona M. Walter, Richard D Neal, Kevin Lybarger, Caroline A. Thompson, Morhaf Al Achkar, Elizabeth A Sarma, Grace Turner, Farhood Farjah, and Matthew J. Thompson. Symptoms and signs of lung cancer prior to diagnosis: case–control study using electronic health records from ambulatory care within a large us-based tertiary care centre. *BMJ Open*, 13, 2023. URL https://api.semanticscholar.org/CorpusID:258240226.

[25] Kory Kreimeyer, Matthew Foster, Abhishek Pandey, Nina Arya, Gwendolyn Halford, Sandra F Jones, Richard Forshee, Mark Walderhaug, and Taxiarchis Botsis. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *J Biomed Inform.*, 73: 14–29, September 2017. ISSN 15320464. doi: 10.1016/j.jbi.2017.07.012. URL https://linkinghub.elsevier.com/retrieve/pii/S1532046417301685.

[26] Ilkin Bayramli, Victor Castro, Yuval Barak-Corren, Emily M. Madsen, Matthew K. Nock, Jordan W. Smoller, and Ben Y. Reis. Predictive structured–unstructured interactions in EHR models: A case study of suicide prediction. *npj Digit. Med*, 5(1):15, January 2022. ISSN 2398-6352. doi: 10.1038/s41746-022-00558-0. URL https://www.nature.com/articles/s41746-022-00558-0.

[27] Denis Jered McInerney, Borna Dabiri, Anne-Sophie Touret, Geoffrey Young, Jan-Willem van de Meent, and Byron C. Wallace. Query-focused EHR summarization to aid Imaging diagnosis. In *Proceedings of the 5th Machine Learning for Healthcare Conference*. Proceedings of Machine Learning Research, 2020.

[28] Emily Alsentzer and Anne Kim. Extractive Summarization of EHR Discharge Notes. *arXiv preprint*, arXiv:1810.12085, 2018.

[29] Jennifer Liang, Ching-Huei Tsou, and Ananya Poddar. A novel system for extractive clinical note summarization using EHR data. In Anna Rumshisky, Kirk Roberts, Steven Bethard, and Tristan Naumann, editors, *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 46–54, Minneapolis, Minnesota, USA, June

2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1906. URL https://aclanthology.org/W19-1906/.

[30] David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. Entity, relation, and event extraction with contextualized span representations. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1585. URL https://aclanthology.org/D19-1585.

[31] Zexuan Zhong and Danqi Chen. A frustratingly easy approach for entity and relation extraction. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.5. URL https://aclanthology.org/2021.naacl-main.5.

[32] Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. Packed levitated marker for entity and relation extraction. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4904–4917, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.337. URL https://aclanthology.org/2022.acl-long.337.

[33] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, June 2019.

Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[34] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234 – 1240, 2020.

[35] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1909. URL https://aclanthology.org/W19-1909.

[36] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1371. URL https://aclanthology.org/D19-1371/.

[37] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1), October 2021. doi: 10.1145/3458754. URL https://doi.org/10.1145/3458754.

[38] Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. Entity-relation extraction as multi-turn question answering. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1340–1350, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1129. URL https://aclanthology.org/P19-1129/.

[39] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21 (140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

[40] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper$_f$*iles/paper*/2020/f*ile*/1457c0d6$bfcb$4967418bfb8ac142f64$a-$*Paper.pdf*.

[41] Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I. Wang, Victor Zhong, Bailin Wang, Chengzu Li, Connor Boyle, Ansong Ni, Ziyu Yao, Dragomir Radev, Caiming Xiong, Lingpeng Kong, Rui Zhang, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. UnifiedSKG: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 602–631, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.39. URL https://aclanthology.org/2022.emnlp-main.39/.

[42] Chia-Hsuan Lee, Hao Cheng, and Mari Ostendorf. Dialogue state tracking with a language model using schema-driven prompting. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages

4937–4949, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.404. URL https://aclanthology.org/2021.emnlp-main.404.

[43] Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, Tao Yu, Noah Smith, and Mari Ostendorf. In-context learning for few-shot dialogue state tracking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2627—-2643. Association for Computational Linguistics, December 2022.

[44] Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. Unified structure generation for universal information extraction. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.395. URL https://aclanthology.org/2022.acl-long.395.

[45] Long Phan, James T. Anibal, Hieu Trung Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. SciFive: a text-to-text transformer model for biomedical literature. *arXiv preprint*, arXiv:2106.03598, 2021.

[46] Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, Emily Alsentzer, et al. Do we still need clinical language models? In *Conference on Health, Inference, and Learning*, pages 578–597. Proceedings of Machine Learning Research, 2023.

[47] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 2022.

[48] Rikhiya Ghosh, Oladimeji Farri, Sanjeev Kumar Karn, Manuela Danu, Ramya Vunikili, and Larisa Micu. RadLing: Towards efficient radiology report understanding. In Sunayana Sitaram, Beata Beigman Klebanov, and Jason D Williams, editors,

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track), pages 640–651, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-industry.61. URL https://aclanthology.org/2023.acl-industry.61.

[49] Yu Gu, Sheng Zhang, Naoto Usuyama, Yonas G. Woldesenbet, Cliff Wong, Praneeth Sanapathi, Mu-Hsin Wei, Naveen Valluri, Erika Strandberg, Tristan Naumann, and Hoifung Poon. Distilling large language models for biomedical knowledge extraction: A case study on adverse drug events. *arXiv preprint*, arXiv:2307.06439, 2023.

[50] Stephen Mayhew, Terra Blevins, Shuheng Liu, Marek Šuppa, Hila Gonen, Joseph Marvin Imperial, Börje F. Karlsson, Peiqin Lin, Nikola Ljubešić, LJ Miranda, Barbara Plank, Arij Riabi, and Yuval Pinter. Universal NER: A gold-standard multilingual named entity recognition benchmark. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4322–4337, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.243. URL https://aclanthology.org/2024.naacl-long.243/.

[51] Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. Biocreative V CDR task corpus: a resource for chemical disease relation extraction. *Database (Oxford)*, 2016.

[52] Xiang Yue, Bernal Jimenez Gutierrez, and Huan Sun. Clinical reading comprehension: A thorough analysis of the emrQA dataset. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4474–4486, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.410. URL https://aclanthology.org/2020.acl-main.410.

[53] Ankit Pal. CLIFT : Analysing natural distribution shift on question answering models

in clinical domain. In *NeurIPS 2022 Workshop on Robustness in Sequence Modeling*, 2022. URL https://openreview.net/forum?id=9PQFROOfqm.

[54] Haoran Zhang, Natalie Dullerud, Laleh Seyyed-Kalantari, Quaid D. Morris, Shalmali Joshi, and Marzyeh Ghassemi. An empirical framework for domain generalization in clinical settings. *Proceedings of the Conference on Health, Inference, and Learning*, 2021.

[55] John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony Beardsworth Costa, Joseph J. Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Medicine*, 15, 2018. URL https://api.semanticscholar.org/CorpusID:49558635.

[56] Bret Nestor, Matthew B. A. McDermott, Willie Boag, Gabriela Berner, Tristan Naumann, Michael C. Hughes, Anna Goldenberg, and Marzyeh Ghassemi. Feature robustness in non-stationary health records: Caveats to deployable model performance in common clinical machine learning tasks. In Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106, pages 381–405. Proceedings of Machine Learning Research, 09–10 Aug 2019. URL https://proceedings.mlr.press/v106/nestor19a.html.

[57] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, and Tao Qin. Generalizing to unseen domains: A survey on domain generalization. In *IJCAI*, 2021.

[58] Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint*, arXiv:2108.13624, 2021.

[59] Alan Ramponi and Barbara Plank. Neural unsupervised domain adaptation in nlp—a survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, 2020.

[60] Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.363. URL https://aclanthology.org/2020.emnlp-main.363/.

[61] Yaroslav Ganin, E. Ustinova, Hana Ajakan, Pascal Germain, H. Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. Domain-adversarial training of neural networks. In *J. Mach. Learn. Res.*, 2016.

[62] Aakanksha Naik and Carolyn Rose. Towards open domain event trigger identification using adversarial domain adaptation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7618–7624, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.681. URL https://aclanthology.org/2020.acl-main.681/.

[63] Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. Adversarial and domain-aware BERT for cross-domain sentiment analysis. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4019–4028, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.370. URL https://aclanthology.org/2020.acl-main.370/.

[64] Sebastian Ruder, Parsa Ghaffari, and John G. Breslin. Data selection strategies for multi-domain sentiment analysis. *arXiv preprint*, arXiv:1702.02426, 2017.

[65] Roee Aharoni and Yoav Goldberg. Unsupervised domain clusters in pretrained language models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online, July 2020. Association

for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.692. URL https://aclanthology.org/2020.acl-main.692/.

[66] David Grangier and Dan Iter. The trade-offs of domain adaptation for neural language models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3802–3813, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.264. URL https://aclanthology.org/2022.acl-long.264/.

[67] Dan Iter and David Grangier. On the complementarity of data selection and fine tuning for domain adaptation. *arXiv preprint*, arXiv:2109.07591, 2021.

[68] Sebastian Ruder and Barbara Plank. Strong baselines for neural semi-supervised learning under domain shift. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1044–1054, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1096. URL https://aclanthology.org/P18-1096/.

[69] Yang Zou, Zhiding Yu, B. V. K. Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, 2018.

[70] Nitay Calderon, Eyal Ben-David, Amir Feder, and Roi Reichart. DoCoGen: Domain counterfactual generation for low resource domain adaptation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7727–7746, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.533. URL https://aclanthology.org/2022.acl-long.533/.

[71] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John Duchi, Vittorio Murino, and

Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems*, NIPS'18, page 5339–5349, Red Hook, NY, USA, 2018. Curran Associates Inc.

[72] Saadullah Amin and Guenter Neumann. T2NER: Transformers based transfer learning framework for named entity recognition. In Dimitra Gkatzia and Djamé Seddah, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 212–220, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-demos.25. URL `https://aclanthology.org/2021.eacl-demos.25/`.

[73] Nanyun Peng and Mark Dredze. Named entity recognition for chinese social media with jointly trained embeddings. In *EMNLP*, 2015.

[74] Thuy-Trang Vu, Dinh Q. Phung, and Gholamreza Haffari. Effective unsupervised domain adaptation with adversarially trained language models. In *EMNLP*, 2020.

[75] Constantinos Karouzos, Georgios Paraskevopoulos, and Alexandros Potamianos. UDALM: Unsupervised domain adaptation through language modeling. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2579–2590, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.203. URL `https://aclanthology.org/2021.naacl-main.203/`.

[76] Fabio Maria Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *CVPR*, 2019.

[77] Zihan Liu, Genta Indra Winata, Peng Xu, and Pascale Fung. Coach: A coarse-to-fine approach for cross-domain slot filling. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of*

the Association for Computational Linguistics, pages 19–25, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.3. URL https://aclanthology.org/2020.acl-main.3/.

[78] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665 – 673, 2020. URL https://api.semanticscholar.org/CorpusID:215786368.

[79] Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In Malvina Nissim, Jonathan Berant, and Alessandro Lenci, editors, *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2023. URL https://aclanthology.org/S18-2023.

[80] Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=gl3D-xY7wLq.

[81] Sang Michael Xie, Ananya Kumar, Robbie Jones, Fereshte Khani, Tengyu Ma, and Percy Liang. In-N-Out: Pre-training and self-training using auxiliary information for out-of-distribution robustness. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jznizqvr15J.

[82] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.378. URL https://aclanthology.org/2023.findings-emnlp.378/.

[83] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=$_V jQlMeSB_J$.

[84] Tushar Khot, Daniel Khashabi, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Text modular networks: Learning to decompose tasks in the language of existing models. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 1264–1279, 2021.

[85] Colin Wei, Sang Michael Xie, and Tengyu Ma. Why do pretrained language models help in downstream tasks? an analysis of head and prompt tuning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 16158–16170. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper$_f iles/paper/2021/file/86b3e165b8154656a71ffe8a327ded7d-Paper.pdf$.

[86] K. Singhal, Shekoofeh Azizi, Tao Tu, Said Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather J. Cole-Lewis, Stephen J. Pfohl, P A Payne, Martin G. Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, P. A. Mansfield, Blaise Agüera y Arcas, Dale R. Webster, Greg S. Corrado, Yossi Matias, Katherine Hui-Ling Chou, Juraj Gottweis, Nenad Tomaev, Yun Liu, Alvin Rajkomar, Joëlle K. Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620:172 – 180, 2022. URL https://api.semanticscholar.org/CorpusID:255124952.

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

[88] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.740. URL https://aclanthology.org/2020.acl-main.740/.

[89] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HklBjCEKvH.

[90] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

[91] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[92] Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Lidén, Zhou Yu, Weizhu Chen, and Jianfeng Gao. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint*, arXiv:2302.12813, 2023.

[93] Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. Generated knowledge prompting for commonsense reasoning. In *Annual Meeting of the Association for Computational Linguistics*, 2021.

[94] Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint*, arXiv:1907.02893, 2019.

[95] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

[96] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard L. Phillips, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In *ICML*, 2021.

[97] Dave Van Veen, Cara Van Uden, Maayane Attias, Anuj Pareek, Christian Bluethgen, Malgorzata Polacin, Wah Chiu, Jean-Benoit Delbrouck, Juan Zambrano Chaves, Curtis Langlotz, Akshay Chaudhari, and John Pauly. RadAdapt: Radiology report summarization via lightweight domain adaptation of large language models. In Dina Demner-Fushman, Sophia Ananiadou, and Kevin Cohen, editors, *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 449–460, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.bionlp-1.42. URL https://aclanthology.org/2023.bionlp-1.42/.

[98] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30(7):1134–1142, 2024. doi: 10.1038/s41591-024-02855-5. URL https://www.nature.com/articles/s41591-024-02855-5.

[99] Fang Zeng, Zhiliang Lyu, Quanzheng Li, and Xiang Li. Enhancing LLMs for impression generation in radiology reports through a multi-agent system. *arXiv preprint*, arXiv:2412.06828, 2024.

[100] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang-Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 12076–12100, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

[101] Joonho Yang, Seunghyun Yoon, ByeongJeong Kim, and Hwanhee Lee. FIZZ: Factual inconsistency detection by zoom-in summary and zoom-out document. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 30–45, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.3. URL https://aclanthology.org/2024.emnlp-main.3/.

[102] A. Chien, H. Tang, B. Jagessar, K. W. Chang, N. Peng, K. Nael, and N. Salamon. AI-assisted summarization of radiologic reports: Evaluating GPT3davinci, BARTcnn, LongT5booksum, LEDbooksum, LEDlegal, and LEDclinical. *AJNR American Journal of Neuroradiology*, 45(2):244–248, February 2024. doi: 10.3174/ajnr.A8102.

[103] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In Sameer Pradhan, Alessandro Moschitti, and Nianwen Xue, editors, *Joint Conference on EMNLP and CoNLL – Shared Task*, pages 1–40, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://aclanthology.org/W12-4501/.

[104] Shany Barhom, Vered Shwartz, Alon Eirew, Michael Bugert, Nils Reimers, and Ido

Dagan. Revisiting joint modeling of cross-document entity and event coreference resolution. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4179–4189, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1409. URL https://aclanthology.org/P19-1409/.

[105] Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew Peters, Arie Cattan, and Ido Dagan. CDLM: Cross-document language modeling. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2648–2662, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.225. URL https://aclanthology.org/2021.findings-emnlp.225/.

[106] Jin Zhao, Nianwen Xue, and Bonan Min. Cross-document event coreference resolution: Instruct humans or instruct GPT? In Jing Jiang, David Reitter, and Shumin Deng, editors, *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 561–574, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.conll-1.38. URL https://aclanthology.org/2023.conll-1.38/.

[107] Qingkai Min, Qipeng Guo, Xiangkun Hu, Songfang Huang, Zheng Zhang, and Yue Zhang. Synergetic event understanding: A collaborative approach to cross-document event coreference resolution with large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2985–3002, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.164. URL https://aclanthology.org/2024.acl-long.164/.

[108] Kawshik Manikantan Sundar, Shubham Toshniwal, Makarand Tapaswi, and Vineet Gandhi. Major entity identification: A generalizable alternative to coreference

resolution. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 11679–11695, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.652. URL https://aclanthology.org/2024.emnlp-main.652/.

[109] Surabhi Datta, Hio Cheng Lam, Atieh Pajouhi, Sunitha Mogalla, and Kirk Roberts. A cross-document coreference dataset for longitudinal tracking across radiology reports. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 3686–3695, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.393/.

[110] Tejas Sudharshan Mathai, Boah Kim, Oana M. Stroie, and Ronald M. Summers. Privacy–preserving large language model for matching findings and tracking interval changes in longitudinal radiology reports. *J Imaging Inform Med.*, 2025. doi: 10.1007/s10278-025-01478-7.

[111] Amos Cahan and James J. Cimino. A learning health care system using computer-aided diagnosis. *Journal of Medical Internet Research*, 19, 2017.

[112] Hagai Rossman, Ayya Keshet, Smadar Shilo, Amir Gavrieli, Tal Bauman, Ori Cohen, Esti Shelly, Ran D. Balicer, Benjamin Geiger, Yuval Dor, and Eran Segal. A framework for identifying regional outbreak and spread of COVID-19 from one-minute population-wide surveys. *Nature Medicine*, pages 1 – 4, 2020.

[113] Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. A general framework for information extraction using dynamic span graphs. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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 3036–3046, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1308. URL https://aclanthology.org/N19-1308/.

[114] Xiangyang Li, Huan Zhang, and Xiao-Hua Zhou. Chinese clinical named entity recognition with variant neural structures based on BERT methods. In *J Biomed Inform.*, page 103422, 2020.

[115] Grace Turner, Kevin Lybarger, Alison Brennan, Elizabeth Chang, Nianiella Dorvall, Jagjeet Gill, Vy Huynh, Kylie Kerker, Jolie Shen, Erica Qiao, Matthew Thompson, and Meliha Yetisgen. Domain adaptation of a deep learning symptom extractor for different patient populations and clinical settings. In *AMIA Annu Symp Abstract*, 2021.

[116] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint*, arXiv:1907.11692, 2019.

[117] Xiaochuang Han and Jacob Eisenstein. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4238–4248, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1433. URL https://aclanthology.org/D19-1433/.

[118] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *EMNLP*, 2018.

[119] Wilson Lau, Laura Aaltonen, Martin L. Gunn, and Meliha Yetisgen-Yildiz. Automatic assignment of radiology examination protocols using pre-trained language models with knowledge distillation. *AMIA Annu Symp Proc.*, 2021.

[120] Andrew Trotman, Antti Puurula, and Blake Burgess. Improvements to BM25 and language models examined. *Proceedings of the 19th Australasian Document Computing Symposium*, 2014.

[121] National Library of Medicine. Overview of SNOMED CT. `https://www.nlm.nih.gov/healthit/snomedct/snomed_overview.html`, Oct 2016.

[122] Sameer Khanna, Adam Dejl, Kibo Yoon, Steven QH Truong, Hanh Duong, Agustina Saenz, and Pranav Rajpurkar. RadGraph2: Modeling disease progression in radiology reports via hierarchical information extraction. In Kaivalya Deshpande, Madalina Fiterau, Shalmali Joshi, Zachary Lipton, Rajesh Ranganath, Iñigo Urteaga, and Serene Yeung, editors, *Proceedings of the 8th Machine Learning for Healthcare Conference*, volume 219, pages 381–402. Proceedings of Machine Learning Research, 11–12 Aug 2023. URL `https://proceedings.mlr.press/v219/khanna23a.html`.

[123] Namu Park, Kevin Lybarger, Giridhar Kaushik Ramachandran, Spencer Lewis, Aashka Damani, Özlem Uzuner, Martin Gunn, and Meliha Yetisgen. A novel corpus of annotated medical imaging reports and information extraction results using BERT-based language models. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1280–1292, Torino, Italia, May 2024. ELRA and ICCL. URL `https://aclanthology.org/2024.lrec-main.115/`.

[124] Sihang Zeng, Yujuan Fu, Sitong Zhou, Zixuan Yu, Lucas Jing Liu, Jun Wen, Matthew Thompson, Ruth Etzioni, and Meliha Yetisgen. Traj-CoA: Patient trajectory modeling via chain-of-agents for lung cancer risk prediction. In *The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance*, 2025. URL `https://openreview.net/forum?id=S4GfRvVTHV`.

[125] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek

Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[126] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[127] Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

[128] Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.findings-emnlp.319`.

[129] Charles S. Dela Cruz, Lynn T. Tanoue, and Richard A. Matthay. Lung cancer: epidemiology, etiology, and prevention. *Clinics in Chest Medicine*, 32(4):605–644, December 2011. ISSN 02725231. doi: 10.1016/j.ccm.2011.09.001.

[130] The National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, 365 (5):395–409, August 2011. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMoa1102873. URL `http://www.nejm.org/doi/10.1056/NEJMoa1102873`.

[131] Harry J. de Koning, Carlijn van der Aalst, Pim A. de Jong, Ernst Th. Scholten, Kris Nackaerts, Marjolein A Heuvelmans, Jan-Willem J Lammers, Carla Weenink, Uraujh Yousaf-Khan, Nanda Horeweg, Susan van 't Westeinde, Mathias Prokop, Willem Mali, Firdaus A A Mohamed Hoesein, Peter M. A. van Ooijen, Joachim G.J.V. Aerts, Michael A. den Bakker, Erik Thunnissen, Johny A Verschakelen, Rozemarijn Vliegenthart, Joan Elias Walter, Kevin ten Haaf, Harry J. M. Groen, and

Matthijs Oudkerk. Reduced lung-cancer mortality with volume CT screening in a randomized trial. *The New England journal of medicine*, 2020. URL https://api.semanticscholar.org/CorpusID:210948029.

[132] Peter Brian Bach, Michael W. Kattan, Mark Thornquist, Mark G Kris, Ramsey C Tate, Matt J. Barnett, Lillian J Hsieh, and Colin B. Begg. Variations in lung cancer risk among smokers. *Journal of the National Cancer Institute*, 95 6:470–8, 2003. URL https://api.semanticscholar.org/CorpusID:15599537.

[133] Adrian Cassidy, Jonathan Myles, Martie van Tongeren, Richard D. Page, Triantafillos Liloglou, Stephen W. Duffy, and John K. Field. The LLP risk model: an individual risk prediction model for lung cancer. *British Journal of Cancer*, 98:270 – 276, 2007. URL https://api.semanticscholar.org/CorpusID:582440.

[134] Martin Carl Tammemägi, Hormuzd A. Katki, William Gray Hocking, Timothy R Church, Neil E. Caporaso, Paul A. Kvale, Anil K. Chaturvedi, Gerard A. Silvestri, Thomas L. Riley, John M. Commins, and Christine D. Berg. Selection criteria for lung-cancer screening. *The New England journal of medicine*, 368 8:728–36, 2013. URL https://api.semanticscholar.org/CorpusID:205094976.

[135] Annette Mcwilliams, Martin C Tammemagi, John R. Mayo, Heidi C. Roberts, Geoffrey Liu, Kam Soghrati, Kazuhiro Yasufuku, S Martel, Francis Laberge, Michel Gingras, Sukhinder Atkar-Khattra, Christine D. Berg, Ken Evans, Richard J. Finley, John Yee, John C. English, Paola Nasute, John R. Goffin, Serge Puksa, Lori Stewart, Scott Tsai, Michael R. Johnston, Daria Manos, Garth Nicholas, Glenwood D. Goss, Jean Morag Seely, Kayvan Amjadi, Alain Tremblay, Paul Burrowes, Paul MacEachern, Rick Bhatia, Ming-Sound Tsao, and Stephen Lam. Probability of cancer in pulmonary nodules detected on first screening CT. *The New England journal of medicine*, 369 10:910–9, 2013. URL https://api.semanticscholar.org/CorpusID:205095145.

[136] Martin C. Tammemagi, Kevin ten Haaf, Iakovos Toumazis, Chung Yin Kong, Summer S. Han, Jihyoun Jeon, John Commins, Thomas Riley, and Rafael Meza. Development and validation of a multivariable lung cancer risk prediction model that includes

low-dose computed tomography screening results. *JAMA Network Open*, 2, 2019. URL https://doi.org/10.1001/jamanetworkopen.2019.0204.

[137] Joan Elias Walter, Marjolein A Heuvelmans, Pim A. de Jong, Rozemarijn Vliegenthart, Peter M. A. van Ooijen, Robin B. Peters, Kevin ten Haaf, Uraujh Yousaf-Khan, Carlijn van der Aalst, Geertruida H. de Bock, W. P. Th. M. Mali, Harry J. M. Groen, Harry J. de Koning, and Matthijs Oudkerk. Occurrence and lung cancer probability of new solid nodules at incidence screening with low-dose CT: analysis of data from the randomised, controlled NELSON trial. *The Lancet. Oncology*, 17 7:907–916, 2016. URL https://api.semanticscholar.org/CorpusID:3550971.

[138] Lung cancer probability in patients with CT-detected pulmonary nodules: a prespecified analysis of data from the NELSON trial of low-dose CT screening. *The Lancet. Oncology*, 15 12:1332–41, 2014. URL https://api.semanticscholar.org/CorpusID:31473401.

[139] Peng Huang, Cheng Ting Lin, Yuliang Li, Martin C Tammemagi, Malcolm V. Brock, Sukhinder Atkar-Khattra, Yanxun Xu, Ping Hu, John R. Mayo, Heidi Schmidt, Michel Gingras, Sergio Pasian, Lori Stewart, Scott Tsai, Jean Morag Seely, Daria Manos, Paul Burrowes, Rick Bhatia, Ming-Sound Tsao, and Stephen Lam. Prediction of lung cancer risk at follow-up screening with low-dose CT: a training and validation study of a deep learning method. *The Lancet. Digital health*, 1 7:e353–e362, 2019. URL https://api.semanticscholar.org/CorpusID:208462551.

[140] Diego Ardila, Atilla P. Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua Jay Reicher, Lily H. Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg S Corrado, David P. Naidich, and Shravya Shetty. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine*, 25:954 – 961, 2019. URL https://api.semanticscholar.org/CorpusID:159041422.

[141] Peter G. Mikhael, Jeremy Wohlwend, Adam Yala, Ludvig Karstens, Justin Xiang, Angelo K Takigami, Patrick P Bourgouin, PuiYee Chan, Sofiane Mrah, Wael Amayri, Yu-Hsiang Juan, Cheng-Ta Yang, Yung-Liang Wan, Gigin Lin, Lecia V.

Sequist, Florian J. Fintelmann, and Regina Barzilay. Sybil: A validated deep learning model to predict future lung cancer risk from a single low-dose chest computed tomography. *Journal of Clinical Oncology*, 41:2191 – 2200, 2023. URL https://api.semanticscholar.org/CorpusID:255773893.

[142] Johannes Brandt, Maulik Chevli, Rickmer Braren, Georgios Kaissis, Philip MÃ¼ller, and Daniel Rueckert. LungEvaty: A scalable, open-source transformer-based deep learning model for lung cancer risk prediction in LDCT screening. arXiv:2511.20116, 2025.

[143] Philip C Prorok, Gerald L. Andriole, Robert S. Bresalier, Saundra S. Buys, David Chia, E. David Crawford, Ronald P. Fogel, Edward P. Gelmann, Fred Gilbert, Marsha A. Hasson, Richard B. Hayes, Christine Cole Johnson, Jack S. Mandel, Albert Oberman, Barbara O'Brien, Martin M. Oken, Sameer Rafla, Douglas J. Reding, W M Rutt, Joel L. Weissfeld, Lance A. Yokochi, and John K. Gohagan. Design of the prostate, lung, colorectal and ovarian (PLCO) cancer screening trial. *Controlled clinical trials*, 21 6 Suppl:273S–309S, 2000. URL https://api.semanticscholar.org/CorpusID:3870509.

[144] Hormuzd A. Katki, Stephanie A. Kovalchik, Christine D. Berg, Li C. Cheung, and Anil K. Chaturvedi. Development and validation of risk models to select ever-smokers for CT lung cancer screening. *JAMA*, 315 21:2300–11, 2016. URL https://doi.org/10.1001/jama.2016.6255.

[145] Hilary A. Robbins, Li C. Cheung, Anil K. Chaturvedi, David R. Baldwin, Christine D. Berg, and Hormuzd A. Katki. Management of lung cancer screening results based on individual prediction of current and future lung cancer risk. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer*, 2021. URL https://api.semanticscholar.org/CorpusID:238989174.

[146] Urmila Chandran, Jenna Reps, Robert Yang, Anil Vachani, Fabien Maldonado, and Iftekhar Kalsekar. Machine learning and real-world data to predict lung cancer risk in routine care. *Cancer Epidemiology, Biomarkers & Prevention*, 32(3):337–343,

March 2023. ISSN 1055-9965, 1538-7755. doi: 10.1158/1055-9965.EPI-22-0873. URL https://aacrjournals.org/cebp/article/32/3/337/718495/Machine-Learning-and-Real-World-Data

[147] Anjun Chen, Erman Wu, Ran Huang, Bairong Shen, Ruobing Han, Jian Wen, Zhiyong Zhang, and Qinghua Li. Development of lung cancer risk prediction machine learning models for equitable learning health system: Retrospective study. *JMIR AI*, 3:e56590, September 2024. ISSN 2817-1705. doi: 10.2196/56590. URL https://ai.jmir.org/2024/1/e56590.

[148] Liwen Sun, Hao-Ren Yao, Gary Gao, Ophir Frieder, and Chenyan Xiong. Intercept cancer: Cancer pre-screening with large scale healthcare foundation models. *CoRR*, abs/2506.00209, June 2025. URL https://doi.org/10.48550/arXiv.2506.00209.

[149] Marvin Chia-Han Yeh, Yu-Hsiang Wang, Hsuan-Chia Yang, Kuan-Jen Bai, Hsiao-Han Wang, and Yu-Chuan Jack Li. Artificial intelligence-based prediction of lung cancer risk using nonimaging electronic medical records: Deep learning approach. *Journal of Medical Internet Research*, 23(8):e26256, August 2021. ISSN 1438-8871. doi: 10.2196/26256. URL https://www.jmir.org/2021/8/e26256.

[150] Lan Wang, Yonghua Yin, Ben Glampson, Robert Peach, Mauricio Barahona, Brendan C. Delaney, and Erik K. Mayer. Transformer-based deep learning model for the diagnosis of suspected lung cancer in primary care based on electronic health record data. *eBioMedicine*, 110:105442, December 2024. ISSN 23523964. doi: 10.1016/j.ebiom.2024.105442. URL https://linkinghub.elsevier.com/retrieve/pii/S235239642400478X.

[151] Iacopo Vagliano, Miguel Rios, Mohanad Abukmeil, Martijn C. Schut, Torec T. Luik, Kristel M. Van Asselt, Henk C. P. M. Van Weert, and Ameen Abu-Hanna. An order-sensitive hierarchical neural model for early lung cancer detection using Dutch primary care notes and structured data. *Cancers (Basel).*, 17(7): 1151, March 2025. ISSN 2072-6694. doi: 10.3390/cancers17071151. URL https://www.mdpi.com/2072-6694/17/7/1151.

[152] Xiudi Li, Erin Y. Yuan, Stephen J. Kuperberg, Clara-Lea Bonzel, Mary I. Jeffway, Tianrun Cai, Katherine P. Liao, Raquel Aguiar-IbÃ¡Ã±ez, Yu-Han Kao, Melissa L. Santorelli, David C. Christiani, Tianxi Cai, and Duan. Early detection of non-small cell lung cancer: an electronic health record data-driven approach. *BMC Med*, 2025. doi: https://doi.org/10.1186/s12916-025-04289-3.

[153] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939785. URL `https://doi.org/10.1145/2939672.2939785`.

[154] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint*, arXiv:1910.01108, 2020.

[155] Spencer C. Evans, Michael C. Roberts, Jared W. Keeley, Jennifer B. Blossom, Christina M. Amaro, Andrea M. Garcia, Cathleen Odar Stough, Kimberly S. Canter, Rebeca Robles, and Geoffrey M. Reed. Vignette methodologies for studying cliniciansâ decision-making: Validity, utility, and application in ICD-11 field studies. *International Journal of Clinical and Health Psychology*, 15(2): 160–170, May 2015. ISSN 16972600. doi: 10.1016/j.ijchp.2014.12.001. URL `https://linkinghub.elsevier.com/retrieve/pii/S1697260014000660`.