MedDec: A Dataset for Extracting Medical Decisions from Discharge Summaries

Mohamed Elgaar[†] **Jiali Cheng**[†] **Nidhi Vakil**[†]
Hadi Amiri[†] **Leo Anthony Celi**[‡]
[†]Miner School of Computer & Information Sciences, University of Massachusetts Lowell
[‡]Institute for Medical Engineering and Science, Massachusetts Institute of Technology
{melgaar,jcheng2,nvakil,hadi}@cs.uml.edu lceli@mit.edu

Abstract

Medical decisions directly impact individuals' health and well-being. Extracting decision spans from clinical notes plays a crucial role in understanding medical decision-making processes. In this paper, we develop a new dataset called "MedDec," which contains clinical notes of eleven different phenotypes (diseases) annotated by ten types of medical decisions. We introduce the task of medical decision extraction, aiming to jointly extract and classify different types of medical decisions within clinical notes. We provide a comprehensive analysis of the dataset, develop a span detection model as a baseline for this task, evaluate recent span detection approaches, and employ a few metrics to measure the complexity of data samples. Our findings shed light on the complexities inherent in clinical decision extraction and enable future work in this area of research. The dataset and code are available through https://github.com/CLU-UML/MedDec.

1 Introduction

Clinical notes contain rich information about medical decision-making. Such notes document patient conditions, medications, laboratory and diagnostic results, assessments and plans, prognoses, and follow-up information, among other crucial data points. However, automatic knowledge extraction from clinical notes has been challenged by imprecise clinical descriptions, heterogeneous data, and the need for data annotation. In particular, although there exist comprehensive and manually verified taxonomies of medical decisions (Braddock et al., 1997; Ofstad et al., 2018), and successful information extraction techniques in medical (Mullenbach et al., 2021; Miwa and Sasaki, 2014; Islamaj Doğan et al., 2011; He et al., 2017; Uzuner et al., 2010), and general (Dethlefs et al., 2012; Goodman, 2002; Frampton et al., 2009; Bui and Peters, 2010; Hsueh and Moore, 2007) domains, there is currently no dataset for extracting (i.e. detecting and classifying) medical decisions in clinical narratives. A medical decision is defined as a particular course of clinically relevant actions and/or a statement concerning the assessment of a patient's health as defined in the Decision Identification and Classification Taxonomy for Use in Medicine (DICTUM) (Ofstad et al., 2016).

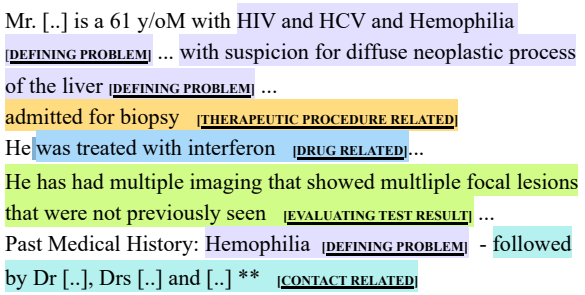

Figure 1: An example excerpt from a de-identified clinical note in MedDec, where text spans are annotated into 10 medical decision categories defined by the Decision Identification and Classification Taxonomy for Use in Medicine (DICTUM) (Ofstad et al., 2016). Color-coded texts represent medical decisions and their annotated decision categories are in [BRACKETS].

Automatic extraction of medical decisions from clinical notes has the potential to transform clinical practice. It can inform the development of evidence-based decision-making guidelines and stewardship programs, identify potential deviations from best decision-making practices, and determine potential risks to patient based on prior medical decisions and their outcomes. In addition, beyond clinical applications, understanding clinical decision patterns can inform health policy development and refinement, especially when evaluating the impact of particular interventions or policies.

This paper develops the first expert-annotated dataset for medical decision extraction and classification within discharge summaries (MedDec). It is developed using patient data sourced from the Medical Information Mart for Intensive Care (MIMIC-

III) (Pollard and Johnson III, 2016), which is a publicly available dataset of de-identified clinical data of patients who were treated in intensive care units (ICUs). MedDec contains annotated decisions in 451 discharge summaries, covering more than 54k sentences and containing diverse patient groups based on *sex*, *race*, and *English proficiency*. In addition, 187 out of the 451 discharge summaries were previously classified into eleven phenotypic (main disease) categories through manual annotation by Gehrmann et al. (2018). We extend the dataset as follows: all medical decisions in the discharge summaries are annotated by domain experts according to the Decision Identification and Classification Taxonomy for Use in Medicine (DICTUM) (Ofstad et al., 2016). DICTUM covers ten (10) medical decision categories listed in Table 1; we add the residual category "None" for spans of texts that do not contain any medical decision. Two expert annotators independently label all text spans of medical decisions in each discharge summary according to the DICTUM guidelines. Disagreements between the annotators are adjudicated by a senior third annotator. Figure 1 shows an excerpt from a de-identified discharge summary annotated with several categories of medical decisions.

In addition, we introduce the new task of *clinical decision extraction*, which involves identifying and classifying spans of medical decisions within relatively long clinical notes. This focused information extraction task contributes to the advancement of bioNLP techniques and has the potential to improve healthcare. We develop several baselines including span detection and named entity recognition models, and evaluate and analyze their performance on MedDec. In addition, we introduce a framework for medical decision extraction to set a baseline for future research.

The contributions of this paper are:

- to the best of our knowledge, MedDec is the first expert-annotated dataset for research on medical decision extraction and classification from clinical notes;
- we provide a comprehensive analysis of the dataset, including distribution reports for protected variables, including sex, race, and English proficiency; and
- we evaluate existing span detection approaches on MedDec, and develop a baseline model to lay the foundation for future research in this area.

2 MedDec

2.1 Taxonomy of Medical Decisions

Table 1 provides descriptions of different types of medical decisions in clinical notes, adapted from DICTUM (Ofstad et al., 2016). The "Contact related" category involves decisions related to patient admissions, discharges, follow-ups, and referrals within the healthcare system. "Gathering information" decisions pertain to acquiring data from sources other than patient interviews or charts, such as ordering tests. "Defining problem" decisions involve complex assessments that define medical issues, including diagnoses, etiological inferences, and prognostic judgments. "Treatment goal" decisions specify treatment objectives beyond general advice. "Drug" decisions pertain to initiation, alteration, cessation, or maintenance of drug regimens. "Therapeutic procedure" decisions involve interventions or therapeutic procedure management. "Evaluating test result" decisions are those that evaluate clinical findings and test outcomes. "Deferment" decisions delay or reject medical decision-making, often due to insufficient information or the need to await test results. "Advice and precaution" decisions involve providing advice or precautions to patients and transferring responsibility for actions to them. Finally, "Legal/insurance-related" decisions deal with medical matters related to legal regulations or financial arrangements.

These categories provide a comprehensive grouping of medical decisions in clinical notes. They can be used for systematic classification and structured analysis of medical decisions, and for understanding the complex processes involved in clinical decision making.

2.2 Data Collection

MedDec is created using patient data sourced from the MIMIC-III (Pollard and Johnson III, 2016). It contains annotated decisions in 451 discharge summaries, representing diverse patient groups based on sex, race, and English proficiency. All medical decisions in each discharge summary are annotated according to the 10 medical decision categories introduced in DICTUM (Ofstad et al., 2016).

The token-level inter-annotator agreement, measured by Cohen's Kappa between the first two annotators is substantial, $k = 0.74$, indicating that it is fairly easy for domain experts to identify medical decisions in discharge summaries. A similar agreement level was reported in DICTUM (Ofstad et al.,

Decision Category	Description	Examples
Contact related	Decision regarding admittance or discharge from hospital, scheduling of control and referral to other parts of the healthcare system	Admit, discharge, follow-up, referral
Gathering information	Decision to obtain information from other sources than patient interview, physical examination and patient chart	Ordering test, consulting colleague, seeking external information
Defining problem	Complex, interpretative assessments that define what the problem is and reflect a medically informed conclusion	Diagnostic conclusion, etiological inference, prognostic judgment
Treatment goal	Decision to set a defined goal for treatment and thereby being more specific than giving advice	Quantitative or qualitative
Drug	Decision to start, refrain from, stop, alter or maintain a drug regimen	Start, stop, alter, maintain, refrain
Therapeutic procedure	Decision to intervene on a medical problem, plan, perform or refrain from therapeutic procedures	Start, stop, alter, maintain, refrain
Evaluating test result	Simple, normative assessments of clinical findings and tests	Positive, negative, ambiguous test results
Deferment	Decision to actively delay a decision or rejection to decide on a problem presented by a patient	Transfer responsibility, wait and see, change subject
Advice and precaution	Decision to give the patient advice or precaution, transferring responsibility for action to the patient	Advice or precaution
Legal/insurance related	Medical decision concerning to legal regulations or financial arrangements	Sick leave, drug refund, insurance, disability

Table 1: Descriptions and high-level examples of medical decisions. The table is re-printed from DICTUM (Ofstad et al., 2016) with slight modification.

2016). We note that token-level agreement provides a lower bound for true inter-annotator agreement as it may sometimes underestimate agreement. This occurs, for instance, when minor variations such as the inclusion or exclusion of less relevant tokens (e.g. stopwords) at the start or end of decision spans are considered as disagreements.

2.3 MedDec Novelty

The novelty of MedDec is in its focused annotation of medical decisions based on an expert-verified and comprehensive taxonomy of medical decisions, its diversity across sex, race, and language proficiency patient groups and phenotypes (diseases), and its potential to drive advancements in both bioNLP research and clinical decision-making. To our knowledge, MedDec is the first dataset specifically developed for extracting medical decisions in clinical notes. This diversity in MedDec enables investigations on potential disparities in medical decisions across the above-mentioned protected variables, which can provide insights for addressing healthcare inequities. These features make MedDec an asset in bioNLP.

2.4 MedDec Statistics

Table 2 reports the percentage of decision spans for each decision category and each protected variable in MedDec. The total counts of decision spans are reported in the last row. Medical decisions related to defining problems, drugs, evaluation, and therapeutic procedures are categories with the highest prevalence, while legal, deferment, treatment goal, gathering information, advice, and contact have considerably lower prevalence. In MedDec, 42.6% of summaries are related to Female patients, 75.9% belong to white patients (of patients with known race), 9.7% to African American, and 85.2% to patients (with known language proficiency) identified as proficient in the English language.

In addition, Table 3 shows the distribution of patients across phenotypic (disease) categories. Patients with psychiatric disorders (including schizophrenia, bipolar, and anxiety disorders), depression, chronic neurologic dystrophies, and chronic pain are more prominently represented in MedDec. Conversely, patients associated with substance abuse, lung conditions, cancer, and obesity are less prevalent in the dataset.

The discharge summaries in MedDec contain 1.4M tokens, with 879K tokens forming part of a span, while 37K tokens belong to more than one span (accounting for 4.2% of labeled tokens). However, the majority of overlaps are minor, where a token marks the end of a span and the start of another.

Decision Type	Sex		Race						Lng. Proficiency	
	Male (n=259)	**Female** (n=192)	**White** (n=327)	**AA** (n=42)	**Hispanic** (n=25)	**Asian** (n=15)	**NH** (n=1)	**Other** (n=21)	**En** (n=260)	**Non-En** (n=45)
Defining Problem	39.2	38.8	39.5	37.5	38.0	36.4	30.9	38.6	38.7	39.2
Drug	26.0	25.1	25.7	24.4	25.0	27.5	19.1	27.0	26.1	25.6
Evaluation	12.9	13.6	12.6	16.6	13.3	12.7	25.5	12.8	13.1	13.9
Therapeutic proc.	12.2	12.4	12.4	12.5	11.7	13.2	10.6	12.2	12.0	12.0
Contact	4.9	5.2	5.0	4.6	6.0	5.4	8.5	4.3	4.8	5.1
Advice	3.4	3.5	3.5	3.2	4.2	3.3	0.0	3.9	3.9	3.0
Gathering info	0.8	0.9	0.8	0.7	1.2	1.3	5.3	0.9	0.9	0.6
Treatment goal	0.3	0.3	0.3	0.3	0.4	0.2	0.0	0.2	0.2	0.4
Deferment	0.2	0.2	0.2	0.2	0.2	0.0	0.0	0.1	0.2	0.2
Legal/Insurance	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Total Count	33,054	24,235	41,666	5,684	3,264	1,737	94	3,078	37,026	6,295

Table 2: Percentage of annotated spans for each decision category across protected variables in MedDec. n is the number of the discharge summaries for each category. The last row shows the total count of decisions per variable.

Decision Types	Phenotypes										
	Substance Abuse (n=8)	**Lung** (n=12)	**Alcohol Abuse** (n=18)	**Psychi-atric** (n=27)	**Obesity** (n=12)	**Heart** (n=23)	**Cancer** (n=12)	**Chronic Neuro** (n=22)	**Depre-ssion** (n=32)	**Chronic Pain** (n=26)	**None** (n=62)
Defining Problem	38.1	36.4	40.5	38.4	37.1	38.9	38.7	40.2	36.8	36.8	38.1
Drug	25.1	32.4	28.5	26.9	29.6	29.1	25.5	26.4	30.3	29.9	24.0
Evaluation	16.2	10.6	11.1	14.0	12.9	11.3	10.4	14.6	13.1	13.2	14.9
Therapeutic proc.	12.8	13.3	11.3	11.9	12.1	12.9	13.7	10.7	12.1	11.4	12.5
Contact	5.5	4.4	4.1	4.9	4.8	3.9	5.2	4.3	4.4	4.7	5.7
Advice	1.1	1.5	2.9	2.7	2.5	2.7	4.7	3.0	2.4	2.8	3.6
Gathering info	0.6	0.9	1.0	0.9	0.3	0.8	1.2	0.5	0.6	0.9	0.7
Treatment goal	0.4	0.3	0.2	0.1	0.3	0.3	0.3	0.1	0.1	0.1	0.3
Deferment	0.2	0.2	0.3	0.2	0.4	0.1	0.2	0.2	0.2	0.2	0.2
Legal/Insurance	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Total Count	2,062	4,319	4,464	8,726	2,957	7,126	2,271	8,301	10,289	8,639	16,790

Table 3: Percentage of annotated spans for each decision category across different phenotypes. n is the number of the discharge summaries for each category. The last row shows the total count of decisions for each phenotype.

3 Learning to Extract Medical Decisions

We present an overview of our approach and describe its two key components in subsequent sections. Figure 2 shows our approach for extracting and classifying medical decisions. First, a *long* note is chunked into segments of acceptable length to the model. Each segment is fed into the model to generate hidden representations and token classification probabilities (this step can be batched for efficiency). Then, the resulting labeled text sequences are concatenated to obtain classification results for the entire clinical note. Finally, we post-process the results to convert them into spans, defined by a start position, end position, and category.

3.1 Sequence Labeling Framework

We develop a multi-class sequence labeling approach that fine-tunes a pre-trained model for span detection. The data consists of a sequence of n tokens $t = \{t_1, \ldots, t_n\}$ and token labels $y = \{y_1, \ldots, y_n\}$, and each label indicates a set of $k+1$ categories indicating k decision types and the none category, $y_i \in \{C_1, .., C_k, O\}$. Practically, the labels follow the BIO (beginning-inside-outside) token labeling scheme (Ramshaw and Marcus, 1995).

We use a pre-trained bidirectional transformer encoder f to encode the tokens and generate d-dimensional hidden states. Formally, the latent representation of the i-th token is computed as:

$$\mathbf{h}_i = f(t)_i, \tag{1}$$

where $\mathbf{h}_i \in \mathbb{R}^d$. We then employ a fully-connected layer $g(.)$ on top of the hidden representation that maps each hidden state to obtain the logits across all classes of decision categories:

$$z_{ik} = g(h_i)_k. \tag{2}$$

To realize multi-class classification, the logits are fed into a softmax function, where the class with the maximum predicted probability is considered the predicted label. This approach does not take

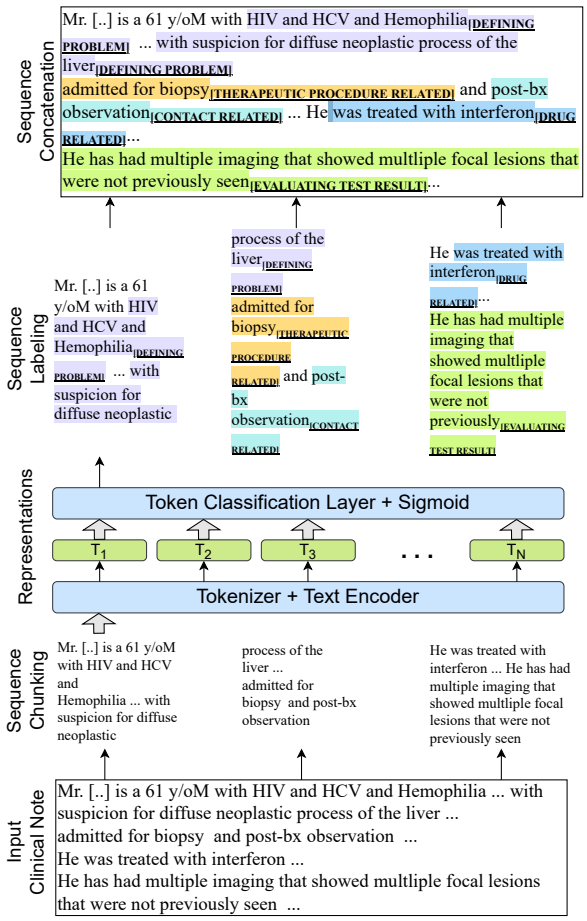

Figure 2: Architecture of the proposed framework for medical span detection. The framework is a multi-class sequence labeling approach that fine-tunes a pre-trained transformer network for span detection.

into account the categorical similarity of neighboring tokens. Previous studies indicate that token classification remains competitive as the most effective span detection and named entity recognition method (Aarsen, 2023; Jurkiewicz et al., 2020; Chhablani et al., 2021; Gu et al., 2022).

There are several alternative methods for span detection, such as detecting the start and end positions, classifying a pair of tokens to check if they constitute a span's boundary, detecting each span category using a separate query process (Devlin et al., 2019; Shen et al., 2022), and conditional random fields (Panchendrarajan and Amaresan, 2018) that compute the maximum probable span assignment from logits.[1]

3.2 Sequence Chunking

Clinical notes are typically thousands of tokens long, and transformer models are computationally restricted and can typically process a maximum of 512 tokens at once. To overcome this challenge, we develop a data sampling function that samples segments of 512 tokens or fewer from random starting points at each training iteration. Therefore, a unique set of text segments is seen at each iteration. This sampling method acts as a data augmentation method by sampling different segments of the same clinical note with different start and end positions, different sentence compositions, and different lengths. At inference time, the input is chunked into segments of 512 tokens with no overlap, each segment is tagged, and the results are concatenated.

[1]Our experiments show that CRF does not improve performance; so, we do not include it in our architecture.

4 Experiments

4.1 Experimental Setup

Models We evaluate the following models:

- **Binder** (Zhang et al., 2023): employs encoders for tokens and token types, optimizes a contrastive objective, with a dynamic threshold loss for negative sampling.
- **PIQN** (Shen et al., 2022): uses NER pointer mechanism (Yang and Tu, 2022) for span boundary detection and an entity classifier for classification. A dynamic label assignment objective is proposed to assign gold labels to instance queries. It dynamically learns query semantics for instance queries and extracts all types of entities simultaneously.
- **DyLex** (Wang et al., 2021): a sequence labeling-based approach that incorporates lexical knowledge with an efficient matching algorithm to generate word-agnostic tag embeddings for NER.
- **Instance-based NER** (Ouchi et al., 2020): formulates NER as instance-based learning, where model assigns labels based on a nearest-neighbor approach.

The following BERT-based models employ the token classification approach described in Devlin et al. (2019). All experiments use the base-size version of the models.

- **DeBERTa v3** (He et al., 2022): uses advanced training strategies, primarily disentangled attention and mask decoder.
- **ALBERT** (Lan et al., 2020): implements shared weights across layers, leading to a greatly reduced memory footprint.

Model	Token Level (Accuracy)	Span Level (F1)	CR (F1)	GI (F1)	DP (F1)	TG (F1)	Dr (F1)	TP (F1)	ETR (F1)	De (F1)	A&P (F1)
ELECTRA	78.2	34.7	**19.9**	0.0	37.9	0.0	47.4	25.2	19.7	**15.4**	35.1
BioClinicalBERT	77.8	34.5	15.9	4.2	38.9	**11.8**	46.4	27.0	19.4	0.0	33.8
RoBERTa	**79.9**	**34.8**	19.3	**5.1**	37.3	6.1	44.7	27.9	23.4	12.5	**42.6**
DeBERTa v3	77.4	31.9	15.2	2.2	32.7	7.4	46.8	24.6	18.5	0.0	28.0
ALBERT v2	74.6	27.8	10.9	4.1	33.0	0.0	38.8	16.6	15.2	0.0	12.0
BINDER	71.2	30.3	17.4	2.5	**59.6**	1.0	**50.9**	**36.8**	**34.0**	0.9	10.2
PIQN	69.5	28.9	16.9	2.4	57.6	1.0	48.9	33.8	32.7	0.9	9.1
DyLex	67.7	27.8	17.4	2.4	57.1	0.9	46.0	31.7	30.1	0.8	0.9
Instance-based	66.2	27.0	16.1	2.5	56.7	0.9	44.3	31.8	28.7	0.8	8.4

Table 4: Span detection performance of different models on MedDec. Span level evaluates the exact match at the span level, while token level evaluates the prediction of decision categories for individual tokens in inputs. *Columns 4-11 show the performance on each decision category, abbreviated according to the order in Table 1*

- **ELECTRA** (Clark et al., 2020): a bidirectional encoder employing a new pre-training objective. It learns to discriminate between real and fake (but plausible) input tokens.
- **RoBERTa** (Liu et al., 2019): a BERT-based model with an improved training objective, hyperparameters, and increased data.
- **BioClinicalBERT** (Alsentzer et al., 2019): a BERT-based model pre-trained on PubMed abstracts and MIMIC-III clinical notes.

Evaluation We use the standard evaluation metrics for NER, span exact match, and token accuracy.[2] The correctness of a span is determined by an exact match (both boundaries and category). We report results in terms of micro-F1 score. Token-level accuracy is a more flexible metric, allowing partial overlap with the true spans.

Difficulty Score We use span length and number of UMLS medical concepts to divide medical decisions into into three difficulty levels, namely Easy, Medium, and Hard. Example of a span with a single medical concept and low cognitive load: "you will continue taking two antibiotics." Example of a span with a high number of medical concepts (underlined) and high cognitive load: "CT abdomen with intravenous contrast: The heart size is at the upper limits of normal. Dense coronary calcifications are identified. In the lung bases, there is bibasilar atelectasis. There are also chronic pleural inflammatory changes including fat deposition and fibrotic changes, left greater than right. Bilateral small pleural effusions are also identified, right greater than left. No focal pulmonary nodules or opacities are identified in the lung bases." These difficulty scores provide insights into the difficulty of learning medical decisions and can also inform curriculum discovery (Elgaar and Amiri, 2023).

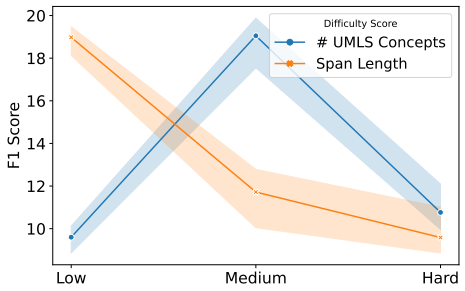

Figure 3: Span detection F1 score on spans with increasing difficulty for two difficulty scores. The shaded area is the 95% confidence interval for three models: ELECTRA, RoBERTa, and BioClinical-BERT.

4.2 Main Results

We compare recent span detection and classification models using our training framework.[3] Table 4 shows the results where we observe that RoBERTa achieves the best performance with a 34.8 F1 score, followed by ELECTRA and BioClinicalBERT.

Span and token accuracy are not perfectly correlated. Although ALBERT performs lower than BINDER in span exact-match, it achieves a higher token accuracy, meaning that it is more effective in partial span detection. BINDER, PIQN, and DyLex achieve higher span-level but lower token-level accuracy than ALBERT, stemming from their design that emphasizes span exact matching.

Our approach focuses on training a model to reliably label segments in clinical notes, irrespective of their boundaries. This strategy results in improved span boundary prediction at inference time.

[2] https://huggingface.co/spaces/evaluate-metric/seqeval

[3] We note that the performance of current models without our training framework is significantly lower.

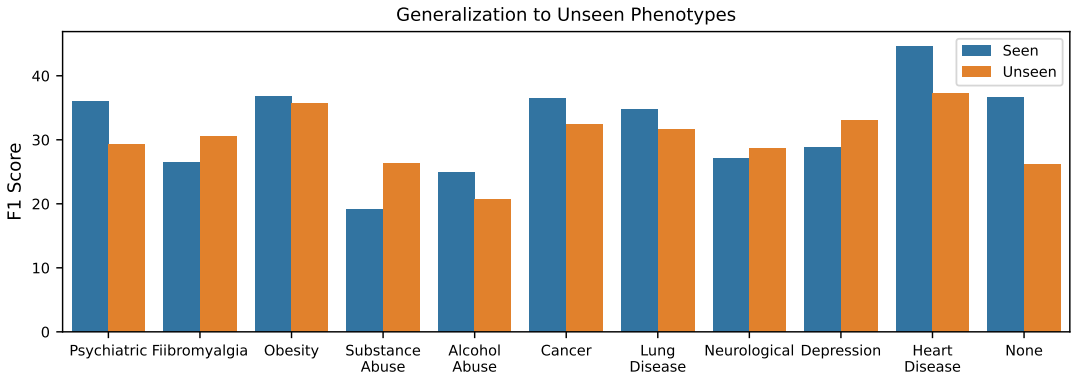

Figure 4: F1 score performance of span detection at phenotype level. The orange bars show the generalizability performance of the model when the phenotype is *unseen* during training.

4.2.1 Span Complexity

We suggest two factors that offer insights into the difficulty of learning spans of medical decisions: (a): the number of medical concepts in each span according to the UMLS ontology, and (b) the length of spans as a heuristic metric; the longer spans may contain more diverse semantic content and can involve more complex sentence structures, which requires the model to maintain more contextual information. We divide the spans into three groups based on their complexity: low, medium, and high.

The results show significant performance disparity across three complexity levels and two metrics. In particular, performance varies considerably based on span length, where predictions are most accurate on shorter spans and least accurate on longer ones. The performance reaches up to 17.7 on easy samples and as low as 10.0 on hard samples. However, we do not observe such a decreasing trend in performance using the number of UMLS concepts, see Figure 3. Span detection performance drops when there is either a very low or very high number of medical concepts in a span. A span with a low number of medical concepts is likely ambiguous and hard to interpret and classify, while a span with a large number of medical concepts is complex and hard to understand. Therefore, sentence length is a better measure of sample complexity, perhaps due to the use of broader contextual information.

We note that the difficulty analysis is the average of the following three models, ELECTRA (Clark et al., 2020), RoBERTa (Liu et al., 2019), and BioClinicalBERT (Alsentzer et al., 2019). Comparing model performance across varying levels of sample complexity provides a better understanding of each model's strengths and weaknesses.

4.2.2 Insights Across Patient Phenotypes

Figure 4 shows a significant disparity in the performance of span detection at the phenotype level, with a standard deviation of 7.5. The highest F1 score is achieved for "Heart Disease" (44.7), while the lowest is for "Substance Abuse" (19.1). Across 11 phenotypes, performance exceeds the average performance (from Table 4) for six phenotypes but falls under the average for the remaining five phenotypes: "Depression," "Neurological Dystrophies," "Substance Abuse," "Fibromyalgia" and "Alcohol Abuse". The variability in the performance of decision extraction across different patient groups identified by different phenotypes highlights the challenges associated with the practical use of the system. We attribute this variance in performance to two potential factors: (a): MedDec contains more training data for the high-performing phenotypes and less training data for the low-performing phenotypes, (b): the intrinsic characteristics of phenotypes affect their difficulty in learning. For example, "Heart Disease" is identifiable through clear clinical markers like abnormal ECG findings or chest pain. Conversely, "Fibromyalgia" is a more complex condition due to its complex nature and subjective symptoms like widespread pain and fatigue. The subjective nature of these symptoms and their overlap with other conditions make it challenging to precisely classify "Fibromyalgia" cases.

4.2.3 Generalizability to Unseen Phenotypes

Results in Figure 4 show a performance drop for 7/11 phenotypes when they are not present in the training data. Conversely, Substance Abuse, Fibromyalgia, Depression, and Neurological do not suffer from being unseen, as perhaps they are sufficiently informed by transferred knowledge from available phenotypes in the training data. Heart

Method	F1 (exact match)	F1 (fuzzy match)
Zero-shot	3.8	10.4
One-shot	4.8	17.9

Table 5: Preliminary analysis of span extraction performance for a prompted LLM in terms of F1 scores of exact and fuzzy match on 10 discharge summaries.

disease and Psychiatric disorders are among the most affected phenotypes, perhaps due to specific domain knowledge related to their decisions that differentiate them from others.

4.2.4 IFT for Span Extraction

Large language models (LLMs) have shown effectiveness in performing a wide variety of tasks through instruction-tuning (IFT) (Zhao et al., 2023; Li et al., 2024; Tran et al., 2024). In tasks such as medical question answering, recent works have shown that LLMs show comparable performance to extensively fine-tuned models (Nori et al., 2023a; Singhal et al., 2023; Thirunavukarasu et al., 2023) using domain-specific prompting methods, such as the retrieval of relevant medical queries to serve as demonstrations (Nori et al., 2023b).

We evaluate `LLama-3-8B-Instruct` (AI@Meta, 2024) on 10 discharge summaries. We prompt the LLM to extract decision spans for each decision category separately, prompting it ten times for each clinical note. We experiment with the zero-shot and one-shot settings both using the following prompt:

```
[[[System]]]
Extract all sub-strings from the
following clinical note that contains
medical decisions within
the specified category.
Print each sub-string in a new line.
If no such sub-string exists, output "None".
[Clinical Note]: {Discharge summary here}

# IF: one-shot setting
[[[User]]]
[Category]: {Decision category here}

[[[Assistant]]]
{Demonstrations}
# End IF

[[[User]]]
[Category]: {Decision category here}

[[[Assistant]]]
{Response}
```

In the one-shot setting, we present as demonstrations all decisions of a single category other than the one being asked for. The demonstration category is selected as the category with the most number of decisions in the clinical note. The results are shown in Table 5. The LLM returns the extracted spans without token-level annotations, therefore it is not possible to calculate token-level accuracy. We compute the performance of span exact match and fuzzy match F1-score. The span fuzzy match is a substitute for token-level accuracy used in Table 4, as the span may be partially detected by the LLM but not be accounted for by the exact match score. To compute the fuzzy span match, we check if either of the extracted span and the true span are a substring of the other, and that they differ by no more than 10 words. These preliminary results show that decision extraction might be challenging for LLMs compared to fine-tuned models.

The low performance of the IFT models can be attributed to the lower efficacy of LLMs in processing long contexts (An et al., 2023). Moreover, the output of LLMs is in free form, which can result in correct responses that do not precisely match the content of medical decisions in notes. For example, a documented decision can be "the patient has high blood pressure," whereas the generated decision can be "the patient is experiencing elevated blood pressure." While semantically correct and relevant, such responses make accurate evaluation of LLM responses a challenging task.

5 Related Work

5.1 BioNLP Datasets

Nye et al. (2018) developed a dataset of 5K abstracts annotated with {*population, interventions, compared, outcomes*} (PICO), to inform personalize patient care. Lehman et al. (2019) developed a dataset of intervention, comparator, and outcome labels of more than 10K randomized controlled trial articles. Patel et al. (2018) developed the clinical entity recognition (CER) corpus, which consists of 5.1K clinical notes annotated by experts with entities such as anatomical structures, body functions, lab devices and medical problems. , and findings. The extracted concepts correspond to a selected group of UMLS semantic types. CLIP (Mullenbach et al., 2021) is a dataset of 718 discharge summaries from MIMIC-III, annotated with seven types of action items: {*Appointment, Lab, Procedure, Medication, Imaging, Patient Instructions,*

Other} at sentence level. covering more than 107K sentences. Stupp et al. (2022) introduced a dataset of 579 admission and progress notes from MIMIC-III, annotated with diseases, assessments, and categories of action items. PHEE (Sun et al., 2022) consists of 5K events from case reports and literature, annotated for pharmacovigilance {*Subject, Drug, Effect*} for drug safety monitoring. Recently, Cheng et al. (2023) developed MDACE, a dataset of clinical notes annotated with ICD codes and their rationales for computer-assisted coding.

5.2 NER and Span Detection

Existing NER and span detection approaches can be divided into sequence labeling-based (tagging) approaches and span-based approaches. Sequence labelling approaches (Aarsen, 2023; Tjong Kim Sang, 2002; Gu et al., 2022) classify every token in the sequence to their corresponding class(es). This formulation is challenged by nested entities. Span-based approaches (Sohrab and Miwa, 2018; Luan et al., 2019; Zheng et al., 2019; Tan et al., 2020; Shen et al., 2021), however, identify and classify spans. First, the spans are either extracted through enumeration or boundary identification and then classified to their corresponding classes.

Du et al. (2019) developed the relational span-attribute tagging (R-SAT) model for extracting clinical entities, their properties, and relations. It employed a method similar to ours, however, the tasks are different as we tackle medical decisions. Ouchi et al. (2020) formulated the NER task as an instance-based learning task, where the NER model was trained to learn the similarity between spans of the same class. DyLex (Wang et al., 2021) retrieved lexicon knowledge for input sequence, applied a denoising module to remove noisy matches, and encoded and fused the lexicon knowledge into the sequence embeddings with column-wise attention for NER.

Abaho et al. (2021) jointly detected and classified spans of health outcomes in clinical notes. Most prior methods decoupled the detection and classification phases. Sent2Span (Liu et al., 2021) was developed for the extraction of PICO information from clinical trial reports. It is designed to work with non-expert sentence-level annotations on the presence of PICO information, without the need of expert span-level annotations and is able to achieve higher recall than comparable methods. PIQN (Shen et al., 2022) developed entity pointer for localization and entity classifier for classification. A dynamic label assignment objective was proposed to extract different types of entities simultaneously.

Recently, Zhang et al. (2023) proposed Binder, which optimized two encoders for NER, one for tokens and one for token types. For each span associated with a class, Binder sampled negative spans based on their loss, and optimized model parameters using a contrastive learning objective. Mirror (Zhu et al., 2023) was an information extraction framework based on graph decoding, where entities were nodes in the graph and the relations of interest were edges. Mirror allowed for extracting all entities and relations in a single step. DICE (Ma et al., 2023) adapted sequence-to-sequence models for structured event extraction from clinical text, using PubMed documents in MACCROBAT dataset (Caufield et al., 2019). Raza and Schwartz (2023) introduced a model consisting of BioBERT (Lee et al., 2019), and Bi-LSTM (Huang et al., 2015) and Conditional Random Field (CRF) (Lafferty et al., 2001) layers to extract clinical (diseases, conditions, symptoms, and drugs) and non-clinical (social factors) entities from clinical notes.

6 Conclusion

We developed MedDec for the extraction and classification of ten types of documented medical decisions in discharge summaries of eleven different phenotypes (diseases). We demonstrate several baseline models to tackle this task. Through extensive experiments and analysis, we find that the task is challenging, the performance of the best-performing model significantly varies across phenotypes and the spectrum of sample complexity. The dataset can be useful in studying population statistics, biases in medical treatment, analysis of medical decisions for different phenotypes, and understanding medical decision-making processes.

Limitations

The present work has several limitations, which form the basis of our future work: the distribution of the ten classes of medical decisions within the dataset is considerably imbalanced. Table 2 highlights these data imbalances across protected variables. Class imbalance may lead to biases in the model training process and affect the model's ability to accurately predict less represented classes. Addressing these data imbalances can prevent com-

putational models from learning and perpetuating such biases in the data. In addition, we note that, these imbalances reflect the challenges of working with real-world data and can inform future research in healthcare equity and the development of systems that perform well across all patient groups. The models have been applied to the notes in MIMIC-III dataset, and the extent of their generalizability to other datasets has not been evaluated. While our best classifier is performant, it may fail to identify and classify certain medical decisions, such as those pertinent to deferment. This limitation could be partly due to the effect of longer decisions, which can inversely affect the model's performance due to the potentially higher linguistic complexity of longer texts (see relevant results in Figure 3). Finally, while discharge summaries contain rich information about patient care, it's important to acknowledge their limitations in fully capturing the breadth of medical decisions made during a patient's hospital stay. Nevertheless, common medical decision-making patterns and clinical reasoning processes are expected to make models trained on these summaries generalize to other types of clinical documents.

Ethic and Broader Impact Statements

This project adheres to ethical considerations and safeguards to ensure the responsible and ethical handling of medical data and its implications. All results have been presented in aggregate, and we have made and will make every effort to protect human subject information and minimize the potential risk of loss of patient privacy and confidentiality (all authors with access to the data have successfully completed a training program in the protection of human subjects and privacy protection). In addition, our work is transformational in nature, and its broader impacts are first and foremost the potential to improve the well-being of individual patients in the society and support clinicians in their medical decision-making efforts.

References

Tom Aarsen. 2023. Spanmarker for named entity recognition. Master's thesis, Radboud University Faculty of Science. Available at https://www.ru.nl/publish/pages/769526/tom_aarsen.pdf.

Micheal Abaho, Danushka Bollegala, Paula Williamson, and Susanna Dodd. 2021. Detect and classify – joint span detection and classification for health outcomes. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8709–8721, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

AI@Meta. 2024. Llama 3 model card. Github repository, accessed on June 6, 2024.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Chenxin An, Shansan Gong, Ming Zhong, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2023. L-eval: Instituting standardized evaluation for long context language models. *ArXiv preprint*, abs/2307.11088.

Clarence H Braddock, Stephan D Fihn, Wendy Levinson, Albert R Jonsen, and Robert A Pearlman. 1997. How doctors and patients discuss routine clinical decisions: informed decision making in the outpatient setting. *Journal of general internal medicine*, 12(6):339–345.

Trung H. Bui and Stanley Peters. 2010. Decision detection using hierarchical graphical models. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 307–312, Uppsala, Sweden. Association for Computational Linguistics.

J Harry Caufield, Yichao Zhou, Yunsheng Bai, David A Liem, Anders O Garlid, Kai-Wei Chang, Yizhou Sun, Peipei Ping, and Wei Wang. 2019. A comprehensive typing system for information extraction from clinical narratives. *medRxiv*, page 19009118.

Hua Cheng, Rana Jafari, April Russell, Russell Klopfer, Edmond Lu, Benjamin Striner, and Matthew Gormley. 2023. MDACE: MIMIC documents annotated with code evidence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7534–7550, Toronto, Canada. Association for Computational Linguistics.

Gunjan Chhablani, Abheesht Sharma, Harshit Pandey, Yash Bhartia, and Shan Suthaharan. 2021. NLRG at SemEval-2021 task 5: Toxic spans detection leveraging BERT-based token classification and span prediction techniques. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 233–242, Online. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Nina Dethlefs, Helen Hastie, Verena Rieser, and Oliver Lemon. 2012. Optimising incremental dialogue decisions using information density for interactive systems. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 82–93, Jeju Island, Korea. Association for Computational Linguistics.

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

Nan Du, Mingqiu Wang, Linh Tran, Gang Lee, and Izhak Shafran. 2019. Learning to infer entities, properties and their relations from clinical conversations. In *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 4979–4990, Hong Kong, China. Association for Computational Linguistics.

Mohamed Elgaar and Hadi Amiri. 2023. HuCurl: Human-induced curriculum discovery. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1862–1877, Toronto, Canada. Association for Computational Linguistics.

Matthew Frampton, Jia Huang, Trung Bui, and Stanley Peters. 2009. Real-time decision detection in multi-party dialogue. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1133–1141, Singapore. Association for Computational Linguistics.

Sebastian Gehrmann, Franck Dernoncourt, Yeran Li, Eric T Carlson, Joy T Wu, Jonathan Welt, John Foote Jr, Edward T Moseley, David W Grant, Patrick D Tyler, et al. 2018. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PloS one*, 13(2):e0192360.

Joshua Goodman. 2002. An incremental decision list learner. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 17–24. Association for Computational Linguistics.

Weiwei Gu, Boyuan Zheng, Yunmo Chen, Tongfei Chen, and Benjamin Van Durme. 2022. An empirical study on finding spans. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3976–3983, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hua He, Kris Ganjam, Navendu Jain, Jessica Lundin, Ryen White, and Jimmy Lin. 2017. An insight extraction system on BioMedical literature with deep neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2691–2701, Copenhagen, Denmark. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2022. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Pei-Yun Hsueh and Johanna D. Moore. 2007. What decisions have you made?: Automatic decision detection in meeting conversations. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 25–32, Rochester, New York. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *ArXiv preprint*, abs/1508.01991.

Rezarta Islamaj Doğan, Aurélie Névéol, and Zhiyong Lu. 2011. A context-blocks model for identifying clinical relationships in patient records. *BMC bioinformatics*, 12(3):1–11.

Dawid Jurkiewicz, Łukasz Borchmann, Izabela Kosmala, and Filip Graliński. 2020. ApplicaAI at SemEval-2020 task 11: On RoBERTa-CRF, span CLS and whether self-training helps them. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1415–1424, Barcelona (online). International Committee for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289. Morgan Kaufmann.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. 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 3705–3717, Minneapolis, Minnesota. Association for Computational Linguistics.

Rumeng Li, Xun Wang, and Hong Yu. 2024. LlamaCare: An instruction fine-tuned large language model for clinical NLP. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10632–10641, Torino, Italia. ELRA and ICCL.

Shifeng Liu, Yifang Sun, Bing Li, Wei Wang, Florence T. Bourgeois, and Adam G. Dunn. 2021. Sent2Span: Span detection for PICO extraction in the biomedical text without span annotations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1705–1715, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. 2019. A general framework for information extraction using dynamic span graphs. 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 3036–3046, Minneapolis, Minnesota. Association for Computational Linguistics.

Mingyu Derek Ma, Alexander Taylor, Wei Wang, and Nanyun Peng. 2023. DICE: Data-efficient clinical event extraction with generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15898–15917, Toronto, Canada. Association for Computational Linguistics.

Makoto Miwa and Yutaka Sasaki. 2014. Modeling joint entity and relation extraction with table representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1858–1869, Doha, Qatar. Association for Computational Linguistics.

James Mullenbach, Yada Pruksachatkun, Sean Adler, Jennifer Seale, Jordan Swartz, Greg McKelvey, Hui Dai, Yi Yang, and David Sontag. 2021. CLIP: A dataset for extracting action items for physicians from hospital discharge notes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1365–1378, Online. Association for Computational Linguistics.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023a. Capabilities of gpt-4 on medical challenge problems. *ArXiv preprint*, abs/2303.13375.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. 2023b. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *ArXiv preprint*, abs/2311.16452.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Melbourne, Australia. Association for Computational Linguistics.

Eirik H Ofstad, Jan C Frich, Edvin Schei, Richard M Frankel, and Pål Gulbrandsen. 2016. What is a medical decision? a taxonomy based on physician statements in hospital encounters: a qualitative study. *BMJ open*, 6(2):e010098.

Eirik Hugaas Ofstad, Jan C Frich, Edvin Schei, Richard M Frankel, Jūratė Šaltytė Benth, and Pål Gulbrandsen. 2018. Clinical decisions presented to patients in hospital encounters: a cross-sectional study using a novel taxonomy. *BMJ open*, 8(1):e018042.

Hiroki Ouchi, Jun Suzuki, Sosuke Kobayashi, Sho Yokoi, Tatsuki Kuribayashi, Ryuto Konno, and Kentaro Inui. 2020. Instance-based learning of span representations: A case study through named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6452–6459, Online. Association for Computational Linguistics.

Rrubaa Panchendrarajan and Aravindh Amaresan. 2018. Bidirectional LSTM-CRF for named entity recognition. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Pinal Patel, Disha Davey, Vishal Panchal, and Parth Pathak. 2018. Annotation of a large clinical entity corpus. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2033–2042, Brussels, Belgium. Association for Computational Linguistics.

Tom J Pollard and AEW Johnson III. 2016. The mimic iii clinical database, version 1.4. *The MIMIC-III Clinical Database. PhysioNet*.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Shaina Raza and Brian Schwartz. 2023. Constructing a disease database and using natural language processing to capture and standardize free text clinical information. *Scientific Reports*, 13(1):8591.

Yongliang Shen, Xinyin Ma, Yechun Tang, and Weiming Lu. 2021. A trigger-sense memory flow framework for joint entity and relation extraction. In *Proceedings of the Web Conference 2021*, WWW '21, page 1704–1715, New York, NY, USA. Association for Computing Machinery.

Yongliang Shen, Xiaobin Wang, Zeqi Tan, Guangwei Xu, Pengjun Xie, Fei Huang, Weiming Lu, and Yueting Zhuang. 2022. Parallel instance query network for named entity recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 947–961, Dublin, Ireland. Association for Computational Linguistics.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *ArXiv preprint*, abs/2305.09617.

Mohammad Golam Sohrab and Makoto Miwa. 2018. Deep exhaustive model for nested named entity recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2843–2849, Brussels, Belgium. Association for Computational Linguistics.

Doron Stupp, Ronnie Barequet, I-Ching Lee, Eyal Oren, Amir Feder, Ayelet Benjamini, Avinatan Hassidim, Yossi Matias, Eran Ofek, and Alvin Rajkomar. 2022. Structured understanding of assessment and plans in clinical documentation. *medRxiv*, pages 2022–04.

Zhaoyue Sun, Jiazheng Li, Gabriele Pergola, Byron Wallace, Bino John, Nigel Greene, Joseph Kim, and Yulan He. 2022. PHEE: A dataset for pharmacovigilance event extraction from text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5571–5587, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chuanqi Tan, Wei Qiu, Mosha Chen, Rui Wang, and Fei Huang. 2020. Boundary enhanced neural span classification for nested named entity recognition. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9016–9023. AAAI Press.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Hieu Tran, Zhichao Yang, Zonghai Yao, and Hong Yu. 2024. Bioinstruct: Instruction tuning of large language models for biomedical natural language processing. *Journal of the American Medical Informatics Association*, page ocae122.

Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518.

Baojun Wang, Zhao Zhang, Kun Xu, Guang-Yuan Hao, Yuyang Zhang, Lifeng Shang, Linlin Li, Xiao Chen, Xin Jiang, and Qun Liu. 2021. DyLex: Incorporating dynamic lexicons into BERT for sequence labeling. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2679–2693, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Songlin Yang and Kewei Tu. 2022. Bottom-up constituency parsing and nested named entity recognition with pointer networks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2403–2416, Dublin, Ireland. Association for Computational Linguistics.

Sheng Zhang, Hao Cheng, Jianfeng Gao, and Hoifung Poon. 2023. Optimizing bi-encoder for named entity recognition via contrastive learning. In *The Eleventh International Conference on Learning Representations*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *ArXiv preprint*, abs/2303.18223.

Changmeng Zheng, Yi Cai, Jingyun Xu, Ho-fung Leung, and Guandong Xu. 2019. A boundary-aware neural model for nested named entity recognition. In *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 357–366, Hong Kong, China. Association for Computational Linguistics.

Tong Zhu, Junfei Ren, Zijian Yu, Mengsong Wu, Guoliang Zhang, Xiaoye Qu, Wenliang Chen, Zhefeng Wang, Baoxing Huai, and Min Zhang. 2023. Mirror: A universal framework for various information extraction tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8861–8876, Singapore. Association for Computational Linguistics.

Text	Context
. . . during that admission, he coded for pulseless vt and was transferred	The first three words of the "Defining Problem" span are not detected.
chief complaint: shortness of breath.	Model failed to extract "Defining Problem" decision.
. . . were orally administered. The patient demonstrated piecemeal behavior by dividing up boluses into multiple swallows regardless of size of consistency of the bolus. There was subsequent premature spillover into the valleculae.	The first sentence is correctly classified as "Evaluating test results". The second sentence is incorrectly classified as "Defining problem" instead of "Evaluating test results".
Coronary angiography in . . . 90% stenosis after d1. The lcx was totally occluded . . . via left collaterals. The rca had a 90% proximal lesion.	Three separate "Evaluating test results" decisions are detected as one. The decision boundary is incorrectly classified.
He was transferred to [..] and neurosurgery was consulted.	The first segment is correctly classified as "Contact related", the second segment is incorrectly classified as "Contact related" instead of "Gathering additional information".

Table 6: Examples where the model fails to extract medical decisions.

A Examples of Model Predictions

Table 6 shows examples where the model partially or fully fails to capture the underlying medical decision from the clinical note.