The Scope of In-Context Learning for the Extraction of Medical Temporal Constraints

Parker Seegmiller  
*Dartmouth College*  
matthew.p.seegmiller.gr@dartmouth.edu

Joseph Gatto  
*Dartmouth College*  
Joseph.M.Gatto.GR@dartmouth.edu

Madhusudan Basak  
*Dartmouth College*  
Madhusudan.Basak.GR@dartmouth.edu

Diane Cook  
*Washington State University*  
djcook@wsu.edu

Hassan Ghasemzadeh  
*Arizona State University*  
Hassan.Ghasemzadeh@asu.edu

Sarah Preum  
*Dartmouth College*  
Sarah.Masud.Preum@dartmouth.edu

John Stankovic  
*University of Virginia*  
stankovic@cs.virginia.edu

Abstract—Medications often impose temporal constraints on everyday patient activity. Violations of such medical temporal constraints (MTCs) lead to a lack of treatment adherence, in addition to poor health outcomes and increased healthcare expenses. These MTCs are found in drug usage guidelines (DUGs) in both patient education materials and clinical texts. Computationally representing MTCs in DUGs will advance patient-centric healthcare applications by helping to define safe patient activity patterns. We define a novel taxonomy of MTCs found in DUGs and develop a novel context-free grammar (CFG) to computationally represent MTCs from unstructured DUGs. Additionally, we release three new datasets with a combined total of \( N = 836 \) DUGs labeled with normalized MTCs. We develop an in-context learning (ICL) solution for automatically extracting and normalizing MTCs found in DUGs, achieving an average F1 score of 0.62 across all datasets. Finally, we rigorously investigate ICL model performance against a baseline model, across datasets and MTC types, and through in-depth error analysis.

Index Terms—health, natural language processing, information extraction, medication information extraction, temporal information extraction, health nlp application

I. INTRODUCTION

According to the CDC, over 48% of the US population uses at least one prescription medicine, and 24% take three or more [3]. However, only four out of every five new prescriptions are filled, and half of those are administered inappropriately [18]. Non-adherence includes incorrectly taking medication concerning a prescription’s suggested time, dosage, frequency, or duration. Non-adherence also includes the mistiming of medication intake with respect to other activities when medication efficacy is temporally dependent on those activities, e.g. eating, exercising, or sleeping [12], [13], [15], [18]. We refer to any temporal constraints associated with medications as medical temporal constraints (MTCs). Non-adherence to MTCs is linked to higher hospital admission rates, increased morbidity, higher healthcare expenses, poor health outcomes, and even death [5], [9], [10], [12], [15]. The effect of violating MTCs can range from minor discomfort to emergency room visits [21].

MTCs are found in drug usage guidelines (DUGs), or medication guidelines. These textual guidelines appear in both formal patient education materials (e.g., drug labels or public health websites [1], [2]), as well as in clinical texts (e.g., prescriptions and after-visit summaries recorded in electronic health records (EHRs) [4]). The variety of MTC sources calls for a generalizable approach to extract and normalize MTCs from such heterogeneous sources.

Although MTCs are critical for medical safety and treatment adherence, to our knowledge, there is no existing solution to formulate and model patient-centric MTCs. This requires (i) creating a flexible and robust computational representation of MTCs, (ii) a dataset of natural language descriptions of MTCs annotated with their computational representations, and (iii) a generalizable solution for mapping descriptions of MTCs to their corresponding computational representations. Addressing these challenges can enhance intelligent systems that improve medication adherence and patient safety [15], [24], [26], [27] or text-based solutions to recommend safe, personalized health information [22], [23].

We formulate and model MTCs for treatment adherence and health safety, in addition to benchmarking the task of extracting MTCs from DUGs. Specifically, (i) we develop a novel taxonomy of potential MTCs and a novel context-free grammar (CFG) based model to represent MTCs from unstructured DUGs computationally. Next, (ii) the taxonomy and CFG are used to label MTCs in three datasets of free-format textual DUGs from heterogeneous sources. Finally, (iii) we define and benchmark the MTC extraction and normalization task using state-of-the-art in-context learning (ICL) strategies, achieving an average F1 score of 0.62 across all datasets. Recent work has demonstrated the generalizability of ICL for extracting health information in the few-shot setting [6], [11], [30]. ICL utilizes a large language model (LLM) to perform a task by
conditioning on a few input-output examples. We also compared ICL to a rule-based baseline model, explore several prompting techniques for ICL, and conduct a thorough error analysis to determine the scope of ICL for this new, safety-critical medical NLP task.

II. MEDICAL TEMPORAL CONSTRAINTS (MTCs)

Modeling MTCs in DUGs is challenging for the following reasons. MTCs vary in terms of temporal precision; some MTCs are definitive, while some are imprecise. Many MTCs constrain a single activity, the medication intake activity (e.g., taking a medication at \( n \) hour intervals). However, MTCs can also form dependencies between multiple activities (e.g., taking a medication \( m \) hours before eating). Based on our review of DUGs from heterogeneous sources [1], [2], [4], [25], we formulate the following novel taxonomy of MTCs.

A. Taxonomy of MTCs

MTCs can be either definitive or imprecise. Definitive MTCs can be further categorized into three classes: dependency, frequency, and interval. Imprecise MTCs can be categorized into four classes: dependency, time dependency, consistency, and time-of-day.

1) Definitive dependency constraints capture temporal dependencies between taking medication and other regular activities. For example, from the DUG for the drug Protonix: “If you are taking the granules, take your dose 30 minutes before a meal.”

2) Frequency constraints capture the temporal constraints regarding the suggested frequency of a medication administration, i.e., how many times a medication should be taken in a specific interval. For example, from the DUG for the drug Wellbutrin: “Take this medication by mouth, with or without food, usually three times daily.”

3) Interval constraints capture the temporal constraints regarding the suggested interval between consecutive medication administrations. For example, from the DUG for the drug Wellbutrin: “It is important to take your doses at least 6 hours apart or as directed by your doctor to decrease your risk of having a seizure.”

4) Imprecise dependency constraints capture inexact temporal dependencies between taking medication and other regular activities. For example, from the DUG for the drug Singular: “Do not take a dose before exercise if you are already taking this medication daily for asthma or allergies. Doing so may increase the risk of side effects.”

5) Time dependency constraints capture inexact temporal dependencies between taking medication and a specific time of day. For example, from the DUG for the medication Prednisone: “If you are prescribed only one dose per day, take it in the morning before 9 AM.”

6) Consistency constraints capture the requirement to take medication consistently at a given time interval. For example, from the DUG for the medication Zocor: “Remember to take it at the same time each day.”

7) Time-of-day constraints capture the requirement to take a medication at a certain time of a day. Take, for example, the DUG for the medication Prednisone: “If you are prescribed only one dose per day, take it in the morning.”

A DUG may contain multiple MTCs for a single medication. For instance, consider the following statement from the DUG of the drug Starlix: “Take this medication by mouth 1-30 minutes before each main meal, usually 3 times daily, or as directed by your doctor.” Here the text has both a definitive dependency constraint (MTC type 1) and a frequency constraint (MTC type 2).

B. A Context-free Grammar for Modeling MTCs

A formal grammar is “context-free” if its production rules can be applied regardless of the context of a nonterminal. The taxonomy of the MTCs mentioned above motivates us to develop a context-free grammar (CFG) to model these definitive and imprecise MTCs. A CFG is a suitable solution to model MTCs as the production rule can be applied to any relevant dataset regardless of the context of the nonterminal, i.e., different types of MTCs. Our novel grammar developed and integrated in this work contains the following set of terminals.

- natural number, \( n \): 1 | 2 | 3...
- activity, \( act \): sleeping | eating | taking medication | ...
- prepositions of temporal dependency, \( dp \): before | after
- prepositions of interval dependency, \( ip \): within | for | apart
- prepositions of occurrence, \( p \): at | in
- unit of time slots, \( u \): hour | minute | day | week
- time stamp, \( t \): the same time | 9 am | 10.30 pm | ...
- time of the day, \( d \): morning | evening | noon

Using these terminals, MTCs can be expressed using the following nonterminals.

1) Definitive dependency constraint: \( V_1 \): \( n.u.dp.act \) (e.g., 30 minutes before eating)

2) Frequency constraint: \( V_2 \): \( n \) times in a \( u \) (e.g., three times a day)

3) Interval constraint: \( V_3 \): \( n.u.ip \) (e.g., 6 hours apart)

4) Imprecise dependency constraint: \( V_4 \): \( dp.act \) (e.g., before meal)

5) Imprecise time dependency constraint: \( V_5 \): \( dp.t \) (e.g., before 9 AM)

6) Consistency constraint: \( V_6 \): \( p.t \) each \( u \) (e.g., at the same time each day or at 9 am each day)

7) Time-of-day constraint: \( V_7 \): \( p.d \) (e.g., in morning)

The proposed CFG can also be used to model compound MTCs. For instance, taking a medication 2 hours before eating \( (V_1) \), 3 times a day \( (V_2) \), and 4 hours apart \( (V_3) \) can be expressed as: \( V_4: V_1.V_2.V_3 \). This grammar can also be extended to model negated MTCs. For instance, “do not take this medication before exercise” can be modeled as, \( \neg V_4 \), where \( V_4: dp.act \).
C. The MTC Extraction Task

We define the task of MTC extraction as an information extraction text-to-structure task, in which a DUG is taken as input and a list of MTCs conforming to the proposed CFG is generated as output. Using a CFG for MTC outputs blends readability and parsability. Consider this statement from the DUG for the drug Pantoprazole, which is used to treat stomach ulcers: “If you are also taking Sucralfate, take Pantoprazole at least 30 minutes before Sucralfate.” The MTC contained in this statement is a definitive dependency constraint (MTC type 1), which under the CFG is labeled “30 minute before taking Sucralfate.” This label is easy to understand while also conforming to the proposed CFG, i.e., \( n = 30 \) \( u = \) minute \( dp = \) before \( act = \) taking Sucralfate. Because it conforms to the CFG, the label enables potential downstream systems to model the semantics of the MTC.

MTCs contained in free-format DUGs may not always conform to the CFG exactly. Consider the statements “take this medication at least 1 hour before any meals” and “be sure to wait at least 1 hour after taking this medication before eating.” Both statements contain the definitive dependency constraint “1 hour before eating” (MTC type 1) however neither statement directly conforms to the CFG. This highlights that grammar-based decoding alone cannot accomplish the task of extracting and normalizing MTCs from these materials. Hence, we examine other methods for extracting MTCs.

D. In-Context Learning for MTC Extraction

Guided by recent breakthroughs in clinical information extraction using LLMs [6], [11], [30], we explore using ICL to benchmark the MTC extraction task. ICL is a recently-introduced paradigm in few-shot sequence-to-sequence text modeling in which an LLM is asked to perform a task after being given a prompt and several examples [6]. We choose to use GPT-3 [8] in our MTC extraction experiments because it is effective in extracting both structured scientific information [11] and medication information, such as dosage and frequency [6], using ICL strategies. Additionally, the number of DUGs across our three datasets is relatively small (< 1000). Lehman et al. show that for size-constrained datasets, ICL with GPT-3 outperforms task-specific models on various clinical tasks [17]. We design two prompts for extracting MTCs from free-format textual DUG data using ICL, and a third model which utilizes customized/specialized prompts for each MTC type.

Simple prompt: This is a simple prompt for extracting all listed MTCs to serve as an ICL baseline.

Guided prompt: The second is a much longer prompt, featuring elements of the labeling guide given to human annotators for annotating the FDA dataset. This prompt includes the rules of the CFG, including lists of both the terminals and the nonterminals. It also includes a list of potential activities. This is referred to as the guided prompt.

Specialized prompt: We also develop prompts for extracting each of the MTC types separately. Each of the prompts contains a basic description of the MTC, as well as a heuristic for formatting the MTC correctly. This approach is referred to as the specialized model.

In addition to each prompt, 20 DUGs are strategically selected from the datasets and passed as few-shot examples, as in [6]. Specifically, these 20 DUGs are selected from all three datasets such that we have a representative sample distribution for each MTC type, i.e., simple and difficult normalization examples, empty and non-empty examples, and examples of single and multiple extracted MTCs. The same 20 examples are used when testing ICL for MTC extraction across all datasets. These examples are removed from the datasets when testing to ensure no data leakage occurs. LLM output is passed through a simple post-processing module which attempts to align the outputs with the CFG. An overview of the ICL system is shown in Fig. 1.

III. DATA

MTCs in DUGs can originate from doctors’ suggestions, patient education materials, or guidelines for prescription medications. We utilize three datasets in this work: the FDA dataset, the Medscape dataset, and the EHR dataset [1], [2], [4]. The first two datasets are derived from patient education materials for prescription medications, while the latter is sourced from prescriptions in EHRs. We use a variety of data sources to demonstrate that our MTC formalization generalizes across DUG domains. The three datasets contain a total of \( N = 836 \) DUGs with labeled MTCs. The labeled datasets are made publicly available to enable future research in MTC extraction1. Examples from each dataset are presented, along with their labeled MTCs, in Table I.

A. FDA Dataset

The openFDA database contains drug product labels for both prescription and over-the-counter drugs submitted to the U.S. Food and Drug Administration (FDA), with text fields such as indications for use, adverse reactions, etc [2]. In this work we utilize the dosage and administration text field, which contains “information about the drug product’s dosage and administration recommendations, including starting dose, dose

1https://zenodo.org/record/7712934#.ZAnurj3MJD9
range, titration regimens, and any other clinically significant information that affects dosing recommendations” [2]. To obtain the FDA dataset, a random sample of 600 drug labels was taken from this database. For each of the 600 drug products sampled, the dosage and administration instructions were annotated for MTCs by two annotators. Using Krippendorff’s alpha coefficient for nominal data, a common measure of inter-annotator agreement for multi-label annotations [16], this annotation resulted in an agreement of 0.74, indicating good agreement. Of the 600 dosage and administration instructions, 371 contained MTCs as defined by our CFG. We refer to these drug product dosage and administration instructions as DUGs, and we refer to these 371 labeled DUGs as the FDA dataset.

B. Medscape Dataset

The Medscape dataset is sourced from 35 real prescriptions of patients with multiple chronic diseases [4], which combined include 83 unique medications. These medications treat several chronic diseases, including but not limited to diabetes mellitus (type I and type II), bipolar affective disorder, depression, hypertension, hypotension, chronic pain, morbid obesity, osteoarthritis, and obstructive sleep apnea. For each of these medications, one or more corresponding DUGs are extracted from a DUG corpus, Medscape [1], [25]. From there, the MTC annotation in the DUG was a three-step process. First, three annotators annotated each sentence in each DUG for whether that sentence contained an MTC or other medical constraints, with 99.4% agreement among all annotators as described in [25]. Second, using a rule base, common temporal phrases were automatically assigned to these DUGs. Finally, a single annotator normalized these automatically extracted phrases to conform to the CFG. It was feasible to assign these MTCs semi-automatically because of recurring lexical patterns of MTCs in the DUG corpus. This process resulted in 121 DUGs, each annotated with one or more normalized MTCs.

C. EHR Dataset

The EHR dataset was extracted automatically from MTSamples, a site containing a large collection of publicly-available, de-identified medical reports submitted by clinics in various medical fields, such as Gastroenterology and Pediatrics [4]. These reports are submitted by many different clinicians, ensuring heterogeneity among extracted DUGs in the EHR dataset. The automatic extraction process involved searching each EHR sample for abbreviated forms of common MTCs. Healthcare professionals use medical abbreviations when writing prescriptions and medical records, some of which directly correspond to MTCs. For example, in this DUG taken from the EHR dataset “the patient has a history of lupus, currently on Plaquenil 200-mg b.i.d.”, the abbreviation “b.i.d.” (Latin “bis in die”) means twice a day. This abbreviation maps to the frequency constraint MTC “two times a day” (MTC type 2). While there are many abbreviations in EHRs, we select 8 that map directly to MTCs. These are listed below with their matching MTCs.

1) **b.i.d.**: 2 times day (MTC type 2)
2) **q.d.**: 1 times day (MTC type 2)
3) **q.h.**: 1 times hour (MTC type 2)
4) **q.i.d.**: 4 times day (MTC type 2)
5) **t.i.d.**: 3 times day (MTC type 2)
6) **h.s.**: before sleep (MTC type 4)
7) **p.c.**: after eating (MTC type 4)
8) **a.c.**: before eating (MTC type 4)

We automatically search through all the EHRs on MTSamples and extract single-sentence statements of appropriate length which include these abbreviations. Using this method, we extract 344 medical report statements and automatically assign MTC labels.

D. Data Characterization

There are 836 labeled DUGs across the three datasets; 371 from the FDA dataset, 121 from the Medscape dataset, and 344 from the EHR dataset. Combined these 836 DUGs contain 1,051 MTCs. The use of three datasets from different sources supports the generalizability of our novel MTC taxonomy. This taxonomy can be used to identify MTCs on both the patient and provider sides since the FDA and Medscape datasets are patient-facing while the EHR dataset is clinician-facing. Statements in the FDA and Medscape datasets typically use the 2nd person perspective when discussing the patient, e.g. “to help you remember, use it at the same time each day.” Statements about patients in the EHR dataset, however, are expressed in 3rd person, e.g. “she was finally put on Effexor 25 mg two tablets h.s.”

While each dataset contains several MTC types, the distribution of MTC types differs across datasets, as seen in Fig. 2. For instance, MTC type 6 is the most common MTC in the Medscape dataset, while it does not appear in the...
FDA dataset. Additionally, the EHR dataset contains almost exclusively frequency MTCs (type 2); 96.51% of MTCs in this dataset are type 2, while the rest are type 4 MTCs. Such idiosyncrasies occur as DUGs from different sources vary with underlying medical conditions and corresponding medications/drugs. This suggests that the CFG is appropriate for multiple types of DUGs. However, MTC-type distributions may vary across DUG domains. While we explore MTCs in drug product dosage and administration labels, prescription drug labels, and de-identified medical records, MTCs may occur in other DUGs such as those found on health education websites, doctor recommendations, and elsewhere.

### Figure 2. Distribution of MTC types across the EHR, FDA, and Medscape datasets.

![Distribution of MTC Types Across Datasets](image)

The ability to computationally represent MTCs is vital for downstream tasks. Hence, MTC labels provided by the annotators conform to the proposed CFG and can be represented symbolically. As an example of a downstream task that utilizes MTC extraction, consider the task of discovering whether a chronic disease patient has violated an MTC of one of their prescription medications. Based on activity patterns recognized by a human activity recognition system, a system could use MTCs extracted from the patient’s prescription medication label to determine whether the patient has violated an MTC, which may lead to poor health outcomes.

### IV. EXPERIMENTAL SETUP

#### A. Classification Metrics

As described in Section II-C, the MTC extraction task is an information extraction text-to-structure task. We choose this task structure to ensure generalizability since there are many possible MTCs according to the proposed CFG and many possible sources of MTCs. For simplicity, however, we evaluate the MTC extraction task as a multiclass multilabel classification task, with each unique extracted MTC treated as a label for a given DUG statement. We consider the union of the MTCs present in the FDA, Medscape, and EHR datasets as the label space, and include an “undefined” label for predicted MTCs that either do not conform to the CFG or do not match any MTCs in the label space. Doing so allows us to use standard multilabel classification evaluation metrics to measure model performance in the MTC extraction task. Unless stated otherwise, we henceforth report the macro average of precision, recall, and F1 [20].

#### B. Validity

In addition to standard classification evaluation metrics, we experiment using a simple heuristic for determining whether an extracted MTC is valid. We define a valid MTC as one which conforms to our proposed CFG, making it able to be represented computationally and thus more useful in downstream tasks. While all invalid extracted MTCs will be incorrect classifications, some incorrectly extracted MTCs will still be valid, indicating that the LLM is learning to format output according to the CFG. Hence, we report the percentage of extracted MTCs that are valid.

#### C. Label Specifics

In the specialized model, since not every DUG contains an MTC of each type, some will have an empty label. We simply insert the label “NONE” for these guidelines to allow the LLM to give a non-empty response. Since there is a large prevalence of empty labels, when investigating the specialized model specifically, we report both the positive class metrics (i.e. macro average metrics across guidelines when excluding guidelines with empty labels) and the macro average metrics across all DUGs.

We note that MTC Type 5 only occurs once across the three datasets, as seen in Fig. 2, and is consequently omitted from experimental results. Although MTC type 5 is discarded from further evaluation, our proposed solutions can be extended to MTC type 5 when there is relevant data.

### V. EXPERIMENTAL RESULTS

We thoroughly examine the scope of ICL for MTC extraction. First, we compare the simple, guided, and specialized ICL prompting strategies for the MTC extraction task. Next we compare ICL performance against a rule-based MTC type classification model. We then further evaluate the specialized ICL model responses, first by dataset and then by MTC type. Finally, we explore the effectiveness of the ICL model responses to extract valid structures from text in the MTC extraction task.

#### A. Prompting Strategies Comparison

Macro average results on the MTC extraction task for the simple and guided prompts, along with the specialized model, are displayed in Table II. While the simple and guided prompts produce poor results overall, the specialized model is able to competently extract MTCs with an F1 score of 0.59. We
hypothesize that extracting MTCs of each type separately, as in the specialized model, allows the LLM to contextualize each MTC type more quickly with fewer examples.

### B. Rule-based Baseline Comparison

To demonstrate the generalizability of the specialized ICL model, we develop a simple rule-based MTC type classification model using common phrases in the Medscape dataset as guidance. As extracting specific MTCs using a simple set of search rules would be difficult, we instead attempt only to identify which MTC types (1-7) occur in a given DUG. We use only the Medscape dataset when developing the rule base, as it has the greatest variety of MTCs, then use the same rule base to identify MTC types across all three datasets. We see in Table III that even when only attempting to identify MTC types in a DUG, a much simpler task than MTC extraction, a rule base developed for the Medscape dataset fails to generalize to either the FDA dataset or the EHR dataset. In comparison, the specialized model generalizes across all 3 datasets in the MTC extraction task. This demonstrates the generalizability of using ICL for the MTC extraction task.

### C. Specialized ICL Results By Dataset

In Table IV we see the results of the specialized model across each of the three datasets. The specialized model performs the best across the EHR dataset, with a macro average F1 score of 0.65. The EHR dataset presents possibly the easiest of the three MTC extraction tasks because all labeled MTCs are mapped one-to-one with medical abbreviations, and there are only two MTC types present in the dataset.

2 As explained in Section III-D, only one imprecise time-of-day dependency MTC (type 5) occurs across the EHR, FDA, and Medscape datasets. Hence, we do not attempt to extract this MTC type.

### D. Specialized ICL Results By MTC Type

In Table V we see the results of the specialized model by MTC type. While the model is able to accurately extract MTCs of most types, it performs best on interval constraints (MTC type 3) with a positive class macro average F1 score of 0.72. The most difficult MTC type for the model to extract is the consistency MTC (type 6), with a positive class macro average F1 score of 0.33. We see that the specialized model frequently hallucinates consistency MTCs and discusses other potential sources of error in Section VI.

### E. Validity

Finally, we explore the ability of the ICL models to produce parsable outputs. We see in Table VI that the specialized model is far more competent at producing parsable output which conforms to the CFG with minimal post-processing, with a 0.99 proportion of valid outputs compared to 0.29 and 0.37 in the simple and guided models, respectively.

### VI. Error Analysis

To investigate model strengths and weaknesses, we sample 60 errors made by the specialized model, 10 of each MTC type. A single human annotator then categorizes each model error, providing one possible reason for each failed MTC

### Table II

**MTC Extraction Results.** We report macro average classification metrics recall, precision, and F1 score.

<table>
<thead>
<tr>
<th>Model</th>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>0.12</td>
<td>0.14</td>
<td>0.12</td>
</tr>
<tr>
<td>Guided</td>
<td>0.15</td>
<td>0.21</td>
<td>0.17</td>
</tr>
<tr>
<td>Specialized</td>
<td>0.57</td>
<td>0.70</td>
<td>0.59</td>
</tr>
</tbody>
</table>

### Table III

**Comparison of Rule-based Model and Specialized Model by Dataset.** The rule-based model predicts which MTC types occur in each drug usage guideline, whereas the specialized model is used for the MTC extraction task. We report macro average F1 across each dataset for both tasks.

<table>
<thead>
<tr>
<th>Model</th>
<th>Medscape Dataset F1</th>
<th>FDA Dataset F1</th>
<th>EHR Dataset F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule-Based Model</td>
<td>0.63</td>
<td>0.43</td>
<td>0.01</td>
</tr>
<tr>
<td>Specialized</td>
<td>0.59</td>
<td>0.61</td>
<td>0.65</td>
</tr>
</tbody>
</table>

### Table IV

**Specialized Model Results by Dataset.** For example, the specialized model is able to extract interval MTCs in the Medscape and FDA datasets with macro F1 scores of 0.50 and 0.70, respectively, whereas no interval MTCs are labeled in the EHR dataset.

<table>
<thead>
<tr>
<th>MTC Type</th>
<th>Medscape Dataset F1</th>
<th>FDA Dataset F1</th>
<th>EHR Dataset F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definitive Dependency (1)</td>
<td>0.65</td>
<td>0.80</td>
<td>–</td>
</tr>
<tr>
<td>Frequency (2)</td>
<td>0.63</td>
<td>0.57</td>
<td>0.79</td>
</tr>
<tr>
<td>Interval (3)</td>
<td>0.50</td>
<td>0.70</td>
<td>–</td>
</tr>
<tr>
<td>Imprecise Dependency (4)</td>
<td>0.45</td>
<td>0.38</td>
<td>0.38</td>
</tr>
<tr>
<td>Consistency (6)</td>
<td>0.63</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Time-of-Day (7)</td>
<td>0.53</td>
<td>0.50</td>
<td>–</td>
</tr>
<tr>
<td>Overall</td>
<td>0.59</td>
<td>0.61</td>
<td>0.65</td>
</tr>
</tbody>
</table>

### Table V

**Specialized Model Results by MTC Type.** We report macro average metrics for both the positive class and overall. We report both the positive class F1 score (i.e., macro average metrics across guidelines when excluding guidelines with empty labels) since there is a large prevalence of empty labels.

<table>
<thead>
<tr>
<th>MTC Type</th>
<th>Positive Recall</th>
<th>Positive Precision</th>
<th>Positive F1</th>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.67</td>
<td>0.65</td>
<td>0.65</td>
<td>0.67</td>
<td>0.70</td>
<td>0.67</td>
</tr>
<tr>
<td>2</td>
<td>0.64</td>
<td>0.57</td>
<td>0.60</td>
<td>0.65</td>
<td>0.66</td>
<td>0.64</td>
</tr>
<tr>
<td>3</td>
<td>0.71</td>
<td>0.74</td>
<td>0.72</td>
<td>0.64</td>
<td>0.81</td>
<td>0.68</td>
</tr>
<tr>
<td>4</td>
<td>0.46</td>
<td>0.46</td>
<td>0.42</td>
<td>0.36</td>
<td>0.49</td>
<td>0.38</td>
</tr>
<tr>
<td>5</td>
<td>0.33</td>
<td>0.32</td>
<td>0.33</td>
<td>0.61</td>
<td>0.65</td>
<td>0.63</td>
</tr>
<tr>
<td>6</td>
<td>0.49</td>
<td>0.50</td>
<td>0.49</td>
<td>0.37</td>
<td>0.66</td>
<td>0.43</td>
</tr>
</tbody>
</table>

### Table VI

**Specialized Model Results by MTC Type.** We report macro average metrics for both the positive class and overall. We report both the positive class F1 score (i.e., macro average metrics across guidelines when excluding guidelines with empty labels) since there is a large prevalence of empty labels.
Hallucinations were especially common in consistency (type 6) and time-of-day (type 7) MTCs, occurring in 44% of labeled MTCs. An example of a consistency (type 6) MTC hallucination is “a checkup every 3 months OR 6 months” interval constraint, as extracted by the specialized model, which was instead labeled as the two semantically-related time-of-day MTCs “every 3 months” and “every 6 months”. Semantically-overlapping errors primarily occurred in 15% of the categorized sample errors.

### VII. RELATED WORKS

#### A. Medical Information Extraction Tasks and Context-Free Grammars

While our work is the first to formalize the MTC extraction task, the broader field of medical information extraction is a vibrant area of research. Most related to the MTC extraction task, much prior work has been done in modeling and extracting temporal and medication information in medical and health texts. A common task is extracting temporal relations between clinical events, such as problems and treatments, in discharge summaries [7], [29]. Another is identifying medication information such as drug names, strengths, and routes in electronic medical records [32], [33]. MTC extraction is related yet novel that is more patient-centric in that it focuses on extracting temporal constraints placed on health-related activities found in DUGs.

Our work additionally focuses on modeling extracted MTCs using context-free grammar. The modeling of temporal phenomena in medical text using CFGs has been leveraged by Hao et al. [19], who introduce a model to leverage CFG to extract temporal expressions in clinical texts. In a similar work, Viani et al. utilize a CFG to parse mental health records and extract the duration of untreated psychosis [31]. Our work is the first to use a CFG to define and extract MTCs in DUGs, and we are the first to experiment with ICL for this task.

#### B. In-Context Learning for Medical Information Extraction

Agrawal, Hegselmann, Lang, Kim, and Sontag have shown that LLMs are able to extract clinical information from the medical text in both the few-shot and zero-shot settings [6]. Specifically, they show that given inputs of clinical discharge summaries or medical abstracts, along with guided prompts, GPT-3 [8] is able to competently perform many medical information extraction tasks such as clinical sense disambiguation, biomedical evidence extraction, and medication extraction. We experiment with similar strategies to benchmark the MTC extraction and normalization task. Related works that utilize ICL...
for structured scientific information extraction include [11], which extracts entities and entity relationships from scientific documents into JSON format, and [30], which formulates the task of extracting social determinants of health from clinical narratives.

VIII. CONCLUSION

In this work, we have developed a novel taxonomy of potential MTCs and a novel CFG based model to computationally represent MTCs found in unstructured DUGs. We present and release three new datasets containing \( N = 836 \) DUGs with labeled normalized MTCs. Finally, guided by recent work in ICL for medical information extraction, we develop and explore an ICL solution for the MTC extraction task, achieving an average F1 score of 0.62 across all datasets. Patient-in-the-loop systems that utilize MTC extraction will have computational representations of patient constraints to guide patient activity, promote medication adherence, and lead to better health outcomes. The taxonomy and CFG of MTCs, dataset of extracted MTCs, and ICL exploration presented in this work will advance patient-centric healthcare applications for treatment adherence.

REFERENCES


Table VIII
EXAMPLES OF IN-CONTEXT LEARNING ERRORS. THE SECOND ERROR, FOR EXAMPLE, IS A HALLUCINATION OF THE CONSISTENCY MTC "SAME TIME EACH DAY" (TYPE 6) GIVEN BY THE SPECIALIZED MODEL WHEN ATTEMPTING TO EXTRACT CONSISTENCY MTCs.

<table>
<thead>
<tr>
<th>Drug Usage Guideline</th>
<th>MTCs (Type Expected)</th>
<th>Model Output</th>
<th>Error Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Take this medication by mouth as directed by your doctor, usually 2 or 3 times daily.</td>
<td>3 times day (2)</td>
<td>2 times day OR 3 times day</td>
<td>Nonvalidity</td>
</tr>
<tr>
<td>One tablet daily or as directed by a physician.</td>
<td>None (6)</td>
<td>Same time each day</td>
<td>Hallucination</td>
</tr>
<tr>
<td>Your doctor may direct you to take it in the morning and at noon.</td>
<td>None (3)</td>
<td>12 hour apart</td>
<td>Semantic Overlap</td>
</tr>
</tbody>
</table>


