

A temporal constraint structure for extracting temporal information from clinical narrative

Li Zhou^a, Genevieve B. Melton^a, Simon Parsons^b, George Hripcsak^{a,*}

^a Department of Biomedical Informatics, Columbia University, 622 West 168th Street, VC5, New York, NY 10032, USA

^b Department of Computer and Information Science, Brooklyn College, Brooklyn, NY, USA

Received 27 May 2005

Available online 29 August 2005

Abstract

Introduction. Time is an essential element in medical data and knowledge which is intrinsically connected with medical reasoning tasks. Many temporal reasoning mechanisms use constraint-based approaches. Our previous research demonstrates that electronic discharge summaries can be modeled as a simple temporal problem (STP).

Objective. To categorize temporal expressions in clinical narrative text and to propose and evaluate a temporal constraint structure designed to model this temporal information and to support the implementation of higher-level temporal reasoning.

Methods. A corpus of 200 random discharge summaries across 18 years was applied in a grounded approach to construct a representation structure. Then, a subset of 100 discharge summaries was used to tally the frequency of each identified time category and the percentage of temporal expressions modeled by the structure. Fifty random expressions were used to assess inter-coder agreement.

Results. Six main categories of temporal expressions were identified. The constructed temporal constraint structure models time over which an event occurs by constraining its starting time and ending time. It includes a set of fields for the endpoint(s) of an event, anchor information, qualitative and metric temporal relations, and vagueness. In 100 discharge summaries, 1961 of 2022 (97%) identified temporal expressions were effectively modeled using the temporal constraint structure. Inter-coder evaluation of 50 expressions yielded exact match in 90%, partial match with trivial differences in 8%, partial match with large differences in 2%, and total mismatch in 0%.

Conclusion. The proposed temporal constraint structure embodies a sufficient and successful implementation method to encode the diversity of temporal information in discharge summaries. Placing data within the structure provides a foundational representation upon which further reasoning, including the addition of domain knowledge and other post-processing to implement an STP, can be accomplished.

© 2005 Elsevier Inc. All rights reserved.

Keywords: Temporal representation; Temporal reasoning; Simple Temporal Problem; Temporal constraints; Temporal model; Natural language processing; Discharge summary

1. Introduction

During the last two decades, researches with different backgrounds, perspectives and objectives have actively conducted research on temporal representation and rea-

soning in medicine. They have attempted to bring together the fundamental methodologies and techniques from different disciplines, including artificial intelligence, database management, and biomedical informatics, to facilitate medical decision making. In general, automated reasoning with temporal data can enhance our understanding of the dynamics of medical phenomena such as symptoms, diagnoses, and interventions and may potentially improve both patient care and medical

* Corresponding author. Fax: +212 305 3302.

E-mail address: hripcsak@columbia.edu (G. Hripcsak).

research. A good amount of work has been published. Two review articles [1,2] present detailed summarization and analysis of this area.

Predominant representations of temporal expressions in the medical informatics literature have consisted of time-date stamps (e.g. 5/7/2003) or durations (e.g. 3 days) stored in clinical databases for the purposes of temporal database storage and retrieval [3–5], temporal abstraction [6–9], and other temporal reasoning systems [1,2]. Important issues such as temporal granularity and indeterminacy [10–12], in addition to other data modeling and processing issues have been studied, e.g. Campbell [13] represented time by time-stamping clinical data with explicit calendar-date and clock-time values.

Effective use of temporal information from narrative clinical notes within the electronic medical record (EMR) represents an imperative challenge for informatics researchers. Representing and reasoning with temporal information contained in narrative data is crucial to the field of natural language processing (NLP) and to the larger field of medical informatics, which will allow NLP output to be modeled temporally with other clinical information, as well as allowing the data to be added to clinical databases and integrated with data from other sources and format. This in turn will allow analysis of medical processes that require temporal reasoning.

Despite recent progress in medical NLP techniques for biomedical applications such as information retrieval, text summarization, and medical error detection [14–22], temporal information from the narrative has not been widely exploited. While some medical NLP systems have represented absolute dates and times and other simple time structures, the capabilities of these systems for temporal reasoning tasks are limited. For example, simple tasks such as detecting events from texts occurring some time before or after another event (e.g. 48 h) are difficult using current NLP techniques [17]. To augment the performance of these systems, it is imperative to achieve improved temporal representations and temporal reasoning with NLP data [1,23–25].

Temporal information in medical text exhibits complex and unique characteristics [24–26]. For instance, relative time is more prevalent compared with absolute time (e.g. *two days ago* rather than *on January 7th, 2004*). In addition, medical events are often temporally related with each other qualitatively (e.g. *fever was before rash*) or quantitatively (e.g. *fever started 3 days before rash*). There are also many medically-specific terms and phrases (e.g. *hospital day #4*). To use temporal information of these types effectively, a significant linguistic analysis must be done in conjunction with the application of medical domain knowledge.

Currently, there is limited work published on representing temporal expressions from medical text. The European Committee for Standardization (CEN) [27] has built a formal temporal representation scheme, but

its main purpose was to build a standard for data exchange amongst healthcare information systems. Johnson [25] proposed a graph-based framework for representing the chronological structure of a clinical text, but the framework assumed that the text had already been processed by an NLP system.

In this study, we built a structure to formally capture the necessary elements of temporal expressions to orient an event on a timeline, specify its duration, and determine the order of an event with respect to other events as they are described in the text. Our method also aimed to accomplish this representation in an expressive, sound, and unambiguous manner. The proposed temporal constraint structure expands the functionality of existing NLP systems for processing temporal information, with the ultimate goal of facilitating higher-level temporal reasoning mechanisms, specifically by allowing for the addition of domain knowledge and integration of NLP extracted data, as well as supporting reasoning about temporal information.

2. Background

Many formalisms for representing and reasoning about temporal knowledge have been proposed using constraint-based approaches [28–33]. In general, these methods use network-based representations, where each node represents a time point and each arc represents temporal constraints which indicate possible temporal relationships between the time points. Each of these methods, however, varies in its expressive power as well as computational tractability. These two characteristics of temporal formalisms tend to be opposing. A method which achieves a suitable tradeoff will be useful for medical informatics applications.

Previously, we demonstrated [24] that among these temporal reasoning mechanisms a Simple Temporal Problem (STP), also formally referred to as a Simple Temporal Constraint Satisfaction Problem (STCSP) [31], was sufficient to represent most temporal assertions in discharge summaries and might be useful for encoding the EMR. An STP is a subset of a Temporal Constraint Satisfaction Problem (TCSP) as proposed by Dechter et al [31]. Unlike a TCSP, an STP does not support temporal disjunctions; that is, it only allows at most one interval constraint on any two time points. In essence, an STP supports metric relations among points, time points anchored in absolute time, and primitive Allen interval relationships [34]; these structures are also efficient and can be solved by polynomial time algorithms [31,35]. We previously modeled electronic discharge summaries as an STP [24]. Each medical event was modeled as an interval with a start and finish. All assertions about events were manually encoded using temporal information in the report, and mapped to the

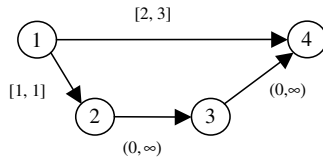


Fig. 1. A simple temporal constraint graph represents Example 2.1. (1, start of headache; 2, start of fever; 3, start of medication; 4, start of admission. Numbers in bracket indicates the temporal constraints between two time points. The finish of each event is equal or after the start of the event by default. The nodes representing the finish of the events are not shown in figure for simplicity.)

model as constraints. A modified all-paired-shortest-paths algorithm was used to derive more information, and the domain was restricted further. Hripcsak et al. [24] contains further details of the modeling process. Fig. 1 depicts a simple temporal constraint graph modeling Example 2.1.

Example 2.1. The patient had headache about 2–3 three days before admission. Fever started one day after headache started. The patient took some medication for the fever before admission.

Though Dechter's paper [31] provided a formal framework for processing temporal constraints, the question of how this representation can be obtained automatically from natural language has not been addressed. Important issues should be studied before applying the STP, including uncertainty, granularity, vagueness, ambiguity, possible sources and solutions for contradiction, as well as implicit constraints based on domain knowledge and linguistic analysis.

Our plan was to build upon our earlier work [24], which proposed that clinical narrative reports can be represented as STPs, with the ultimate objective of developing a comprehensive treatment of temporal information in medical narratives. We presented system architecture in [26], which includes temporal information annotation, extraction, and reasoning by incorporating NLP techniques, domain knowledge, efficient reasoning algorithms, and other applicable methods. Fig. 2 shows how the information flows from the origi-

nal clinical narrative to time-oriented information for medical applications. The formal structure introduced in this paper is an important step towards these goals, as the structure can be used to (1) integrate these expressions with domain knowledge, (2) incorporate context-based knowledge output from the general NLP system, and (3) finally implement these expressions as an STP. In addition, the generalizability of the proposed temporal constraint structure allows the structured information to be applied to other reasoning approaches.

We chose discharge summaries as our study corpus for several reasons. First, discharge summaries are concise clinical reports that describe: (1) the history of present illness leading up to admission; (2) evaluation, progression, and assessment (including procedures, treatments or services provided) of the patient during hospitalization; (3) discharge appraisals and diagnoses; and (4) follow-up plans. Thus, data from discharge summaries represent a comprehensive picture of a patient's hospital course. Second, discharge summaries are importantly often electronically accessible and can be successfully parsed by available NLP systems and used to facilitate different informatics applications.

3. Methods

3.1. Temporal constraint structure

A corpus of 200 random discharge summaries from the Columbia University Medical Center data repository, which covers the period of 1987 to 2004 and containing 134,322 words in total, was used in a bottom-up, grounded approach to create a model framework for medical temporal information encoding. As different expressions were examined, patterns were recognized, and natural categories were identified empirically. To evaluate the diversity of temporal expressions in clinical narratives, the distribution of these categories they fall in, a subset of 100 discharge summaries was used to tally the frequency of each major category of temporal expressions and their subcategories.

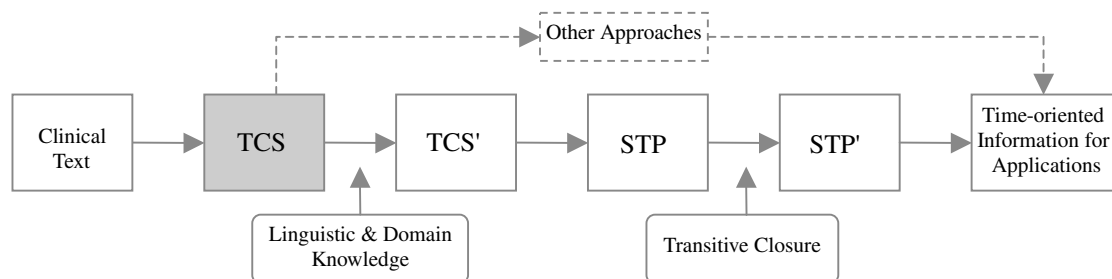


Fig. 2. How the information flows from the original clinical narrative to time-oriented information for applications is shown. TCS, raw temporal assertions; TCS', modified assertions using linguistic and domain knowledge; STP, assertions recast as mathematical constraints; STP', all known temporal relationships. The generalizability of the temporal constraint structure allows the information to be applied to other reasoning approaches.

An iterative process was conducted to develop the model. One author (LZ, an informatics Ph.D. candidate with a medical degree) examined all the temporal expressions in each category, and the initial model was proposed by collaborating with another author (GH, an informatics faculty member and internist). Then the model was presented to the other authors and to a clinical data mining research group for further modification. As such, after several rounds of revision and enhancement, the final formal representation entitled the **temporal constraint structure** was constructed.

We adopted a broad definition of medical events, as previously described [24], where a medical event is defined as a process with a start and finish (e.g. a procedure), the presence of a state for some period of time (e.g. the presence of nausea), or an instantaneous change (e.g. death). Our initial analysis revealed that medical events could be conceptualized as occurring as an interval with two endpoints where the finish point is never before the start point. Furthermore, temporal expressions assert when and how long events occur. Thus, it appeared to be possible to represent temporal expressions by placing an endpoint of the event in a relative time line and determining the relative or metric relationship between the endpoint and the anchor in the expression. The temporal constraint structure was developed based on the above conception. For example, for the statement “*The fever started two days before admission*,” we can infer that “the **start** (an endpoint of an event) of **fever** (a medical event) is **2 days before** (quantity, time unit along with direction) **admission** (an anchor).” The structure contains a set of components in a formal way and represents this information soundly and unambiguously.

3.2. Evaluation of the temporal constraint structure

In the 100 discharge summaries, the percentage of temporal expressions modeled by the temporal constraint structure was calculated. To assess the reliability of coding, a subset of 50 random temporal expressions was encoded by another author (GH). Raw inter-coder agreement was calculated for encoding temporal expressions using the temporal constraint structure. Four criteria of match were defined, including exact match, partial match with trivial differences, partial match with large differences, and total mismatch. Partial match with trivial differences indicates minor differences due to reasons such as naming and typographic error. Partial match with large differences indicates subtle differences due to reasons such as ambiguous context. These differences do not affect the original meaning of the temporal expression upon which the two coders agree with each other. In contrast, total mismatch means that the structures encoded by the two coders for the same expression have totally different meanings.

It is important to note that the expressions that were modeled in the evaluation were historical events (things that are recorded in the discharge summaries as having happened before the summaries were written). Hypothetical events, including medical plans/follow-up plans, therapeutic plans, and prognosis, are beyond the scope of this paper. Though the temporal constraint structure can be applied to hypothetical events, they were modeled as statement of the plan, for example, “thiamine 100 mg p.o. q.d.” was modeled as a statement of a patient’s medication.

4. Results

In this section, we start by examining the diversity of temporal information in clinical narrative data, and then answer the question whether we can model all the temporal expressions in the test set using the temporal constraint structure. We organize most part of this section by temporal categories. This may also help readers to gain a composite view of the modeled data.

Analysis of temporal expressions in the overall corpus of discharge summaries identified six major categories of temporal expressions: “date and time,” “relative date and time,” “duration,” “event-dependent temporal expression,” “fuzzy time,” and “recurring times.” Some of these classes were naturally arranged into subcategories. Two thousand and twenty-two temporal expressions contained in the subset of 100 discharges summaries were manually categorized and the frequency of each category was calculated. The results are shown in Table 1 and exhibit the diversity of temporal expressions in the corpus. Some temporal expressions fell into more than one category. However, for simplification, each expression was classified into only one category according to which category was judged by the coder as the predominant feature of the temporal expression. For example, “*two days before admission*” has a duration, and it is anchored by an event “admission.” Therefore, we classified it as an event-dependent temporal expression. The scope of each category is introduced in Section 4.1.

Each category is discussed in detail in the context of the final constructed model. Most of the temporal expressions were modeled using the temporal constraint structure. However, two special issues need to be addressed here: (1) expressions that were modeled in our evaluation were historical events. Since medication dosing was also considered a medical plan, we folded the medical dosing into the event name so that the frequency information was preserved. Other recurring events were modeled in a similar way. The occurrences of medical dosing varied in discharge summaries. However, to show overall occurrences of other recurring events, we counted them in Table 1. (2) Similar to their treatment

Table 1
The different categories, subcategories of temporal expressions in 100 discharge summaries

Categories	Subcategories	Frequency (%)	Examples
Date and time	Date	731 (36.15)	On 01/25/1991, in May, in 2000 (admission, discharge, dictation, and transcription date contained in each report were also counted)
	Time of day	17 (0.84)	At 5:30, at 4 pm
	Part of day	33 (1.63)	In the morning, at night, am
	Time Range	9 (0.45)	From October 20 to the 31st
	Conjunction	1 (0.05)	Between January 22nd and 28th
	Disjunction	0 (0.00)	On April 4th or 6th
	Others	8 (0.40)	Weekdays, seasons, holidays, etc.
	Subtotal	799 (39.52)	
Relative date and time	Yesterday, today, tomorrow	7 (0.35)	Yesterday, today, tomorrow
	Past/next (time unit)	16 (0.79)	The next day, in the past month
	(A period of time) Ago/after	59 (2.92)	25 years ago, 3 h later
	In/within (a period of time)	28 (1.38)	Within three days, in the last four days
	Subtotal	110 (5.44)	
Duration	Treatment duration	41 (2.03)	A 10 days of antituberculous
	Others	144 (7.12)	For the past 3 years, for 10 min
	Subtotal	185 (9.15)	
<i>Event-dependent temporal expressions</i>			
Key events	Admission	161 (7.96)	One week prior to admission, on admission
	Specific hospital day	19 (0.94)	On the second hospital day
	Hospital stay	38 (1.88)	During the hospitalization, during this admission
	Discharge	56 (2.77)	By the date of discharge
	Operation	96 (4.75)	On post-operative day #1, post-operatively
	Other key events	16 (0.79)	Parturition, transfer
	Previous hospitalization	10 (0.49)	Last hospitalization
	Subtotal	396 (19.58)	
Other events		172 (8.51)	She has noted dizziness since the car accident
	Subtotal	568 (28.09)	
Fuzzy time	Past	175 (8.65)	Status post, history of hypertension, in the past
	Present	42 (2.08)	Now, currently
	Future	3 (0.15)	In the future
	This time/that time	79 (3.91)	At that time, in that same month
	Non-specific time	23 (1.14)	His mood was lowest in the morning
	Subtotal	322 (15.92)	
Recurring time		38 (1.88)	He felt guilty almost everyday (medicine dosing was not counted)
Total		2022 (100.00)	

by others [36], temporal adjectives and adverbs such as “occasional” and “chronic” were also folded into the event names due to their lack of specific meaning and for practical reasons. For example, for “occasional” in “occasional cough for three months,” the event would be “occasional cough” and time duration modeled would be “three months.” These expressions were not counted in Table 1.

From these fundamental principles and empirical observations, we built a representation to incorporate temporal expressions, entitled the **temporal constraint structure**. The proposed temporal constraint structure contains a set of fields that construct a constraint interval(s) and constrain the start and the finish of medical events. Assertions generally specify the constraint interval by stating an explicit time (e.g. a date) or by stating

an anchoring event (e.g. admission), or by applying a temporal operation to an anchor (e.g. two weeks before admission). The definition of each field of the temporal constraint structure and possible values are described in Table 2. The example, “his cough started at least two weeks before admission,” is conceptualized into a time line as shown in Fig. 3. For this temporal assertion, the event (*cough*) itself is an interval with a start and a finish. We construct a constraint interval, which is a semi-bounded interval having an endpoint at two weeks before the *Start of admission* (*Sa*), to constrain the start of the event interval (*Start of cough—Sc*). *Sc* is constrained to occur within the constraint interval. In addition, note that two weeks before *Sa*, while it appears to specify a point, actually generates an interval (or a time range). The author of that phrase does not intend to

Table 2
Fields of the temporal constraint structure

Fields	Definition	Values
<i>event_point^a</i>	Endpoint(s) (the start and/or finish) of the event; constrained by the temporal expression values	<i>Start, finish, both</i> (e.g. “the myocardial infarction occurred on July 5, 2002”), or <i>unspecified</i> (the endpoint of the event being constrained is not explicitly stated in the text)
<i>anchor^a</i>	Constraining time point (e.g. in “he had an operation on October 1st, 2004,” the anchor is 2004-10-01)	A calendar date, a time of day, a relative date or time, an event, or a time reference which may have values of <i>narrative reference, previous reference and now</i> . (e.g. “at that time” can be encoded as <i>previous reference</i>)
<i>anchor_point</i>	If anchor is an event, the endpoint of the event is specified	<i>Start, finish, both, or unspecified</i> , which are similar to <i>event_point</i> (e.g. to encode “after the radiation therapy,” anchor is <i>radiation therapy</i> and the anchor_point is <i>finish</i>)
<i>anchor_modifier</i>	Indicates the stage of a period of time, or the course of an event (e.g. anchor “1980s” in “early 1980s” has a modifier of “early”)	<i>Early, mid, and late</i>
<i>relation^a</i>	A temporal relation between an endpoint of an event and its anchor or interval(s) constructed by the constraint structure with respect to the anchor	<i>Equal, before, equal or before, after and equal or after</i>
<i>time_unit</i>	Unit for measuring time periods	<i>Second, minute, hour, day, week, month, and year</i>
<i>quantity</i>	Specified or indefinite number or amount for measuring the length of a time period	A number, a vague quantifier (e.g. <i>many, a few</i>), or <i>unspecified</i> (e.g. <i>for years</i>)
<i>direction</i>	Indicates the direction of an interval relative to its anchor	<i>Minus</i> (in the past), <i>plus</i> (in the future), or <i>both</i> (e.g. <i>within two days</i>)
<i>interval_operator</i>	Characterizes an endpoint of the event; determines whether an endpoint of an event occurred a specified duration away from the anchor (<i>jump</i>), or any time between the anchor and a specified duration away from the anchor (<i>drag</i>)	<i>Jump</i> (e.g. <i>two days ago</i>), or <i>drag</i> (e.g. <i>within the last two days</i>)
<i>vagueness</i>	Indicates if a vagueness modifier is contained within the expression (e.g. <i>approximately, about, roughly, and more or less</i>)	<i>Yes</i>

^a These fields are required for the temporal constraint structure.

specify a point that is an exact number of seconds before Sa, but rather to provide a range (e.g. two weeks plus or minus some vagueness factor). We initially map the

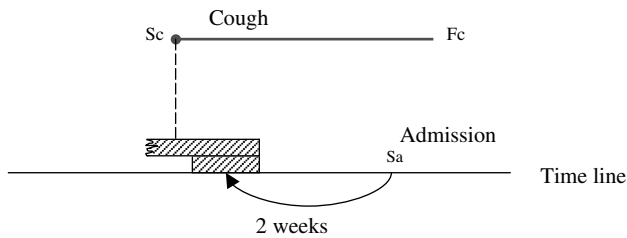


Fig. 3. Example temporal expression “his cough started at least two weeks before admission” modeled on a time line, depicting the event “cough” anchored by “admission.” (Sc, start of cough; Fc, finish of cough; Sa, start of admission; the bars represent the vagueness of the temporal information, which widen the limits of the constraints) The example can be captured using the constraint structure as follows: (note that there is a difference between the constraint structure and the figure about “2 weeks.” In the temporal constraint structure, we used the exact number of time units as it was stated. However, later on in our post-processor, we will convert this exact number to a time range to represent the vagueness of the information, which is shown as a bar) event = “cough”; event_point = “start”; relation = “equal_or_before”; quantity = “2”; time_unit = “week”; direction = “minus”; interval_operator = “jump”; anchor = “admission”; anchor_point = “start”.

phrase to a temporal constraint structure that looks as if it specifies time points to represent what is stated, but then we can use knowledge to generate wider constraint intervals in a post-processing stage.

The details of representing temporal information using the temporal constraint structure are introduced as follows by categories. Table 3 includes several representative examples for each category with accompanying pictorial illustrations.

4.1. Major categories of temporal expressions in modeling with the temporal constraint structure

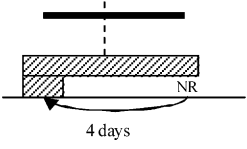
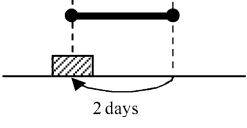
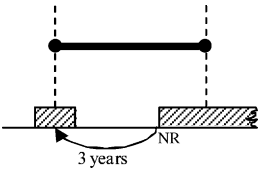
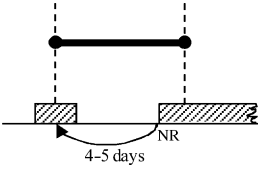
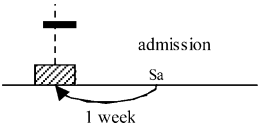
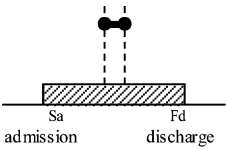
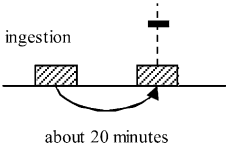
4.1.1. Date and time

Often, temporal expressions are stated in terms of a *calendar dating system* (typically identified by its calendar year/month/date), *date-of-week* (e.g. Monday, Tuesday, and Saturday) *time-of-day system* (expressed by unit of time—hour, minute, and second), *part-of-day* (e.g. morning, evening, daytime, and a.m.), or *four seasons* (e.g. spring, summer, autumn, and winter). To represent these date and time expressions using the constraint structure, the expression is represented with

Table 3
Examples of temporal expressions encoded by the temporal constraint structure with pictorial illustration

Statement	Figure	Temporal constraint structure encoding
4.1.1. Date and Time		
4.1.1.1. <i>He was a heavy smoker until 1984</i>		event = "smoking"; event_point = "finish"; anchor = "1984"; relation = "equal"
4.1.1.2. <i>the myocardial infarction occurred on January 5th,2003</i>		event = "myocardial infarction"; event_point = "both"; anchor = "2003-01-05"; relation = "equal"
4.1.1.3. <i>The patient was treated from October 20th to the 31st with clindamycin.</i>		event = "treat with clindamycin"; event_point1 = "start"; anchor1 = "--10-20"; relation1 = "equal"; event_point2 = "finish"; anchor2 = "--10-31"; relation2 = "equal"
4.1.1.4 <i>He had flu between Jan.3rd and Jan. 10th,2003</i>		event = "flu"; event_point1 = "start"; anchor1 = "2003-01-03"; relation1 = "equal or after"; event_point2 = "finish"; anchor2 = "2003-01-10"; relation2 = "equal or before"
4.1.2. Relative date and time		
4.1.2.1. <i>Her rash started today</i>		event = "rash"; event_point = "start"; anchor = "today"; relation = "equal"
4.1.2.2. <i>He had a laminectomy about 25 years ago</i>		event = "laminectomy"; event_point = "unspecified"; anchor = "narrative_reference"; relation = "equal"; quantity = "25"*; time_unit = "year"; direction = "minus"; interval_operator = "jump"; vagueness = "yes"
4.1.2.3. <i>One month later, he was transfused three units of packed cells.</i>		event = "transfusion"; event_point = "unspecified"; anchor = "narrative_reference"; relation = "equal"; quantity = "1"*; time_unit = "month"; direction = "plus"; interval_operator = "jump"
4.1.2.4. <i>The patient had an orthopedic clinic visit within the last three days.</i>		event = "orthopedic clinic visit"; event_point = "unspecified"; anchor = "narrative_reference"; relation = "equal"; quantity = "3"*; time_unit = "day"; direction = "minus"; interval_operator = "drag"

Table 3 (continued)

Statement	Figure	Temporal constraint structure encoding
4.1.2.5. <i>She has had decreased appetite in the last four days</i>		event = "decreased appetite"; event_point = "unspecified"; anchor = "narrative_reference"; relation = "equal"; quantity = "4"*; time_unit = "day"; direction = "minus"; interval_operator = "drag"
4.1.3. Duration		
4.1.3.1. <i>The pain lasted for 2 days</i>		event = "pain"; event_point = "start"; anchor = "event"; anchor_point = "finish"; relation = "equal"; quantity = "2"*; time_unit = "day"; direction = "minus"; interval_operator = "jump"
4.1.3.2. <i>He has a history of hypertension for the past 3 years.</i>		event = "hypertension"; event_point1 = "start"; anchor1 = "narrative_reference"; relation1 = "equal"; quantity1 = "3"*; time_unit1 = "year"; direction1 = "minus"; interval_operator1 = "jump"; event_point2 = "finish"; anchor2 = "narrative_reference"; relation2 = "equal or after"
4.1.3.3. <i>She had upper respiratory infection for the past 4 to 5 days</i>		event = "upper respiratory infection"; event_point1 = "start"; anchor1 = "narrative_reference"; relation1 = "equal"; quantity1A = "4"*; time_unit1A = "day"; quantity1B = "5"*; time_unit1B = "day"; direction1 = "minus"; interval_operator1 = "jump"; event_point2 = "finish"; anchor2 = "narrative_reference"; relation2 = "equal or after"
4.1.4. Event-dependent temporal expression		
4.1.4.1. <i>He had a fever one week prior to admission</i>		event = "fever"; event_point = "unspecified"; anchor = "admission"; anchor_point = "start"; relation = "equal"; quantity = "1"*; time_unit = "week"; direction = "minus"; interval_operator = "jump"
4.1.4.2. <i>During this hospitalization, he had a blood transfusion.</i>		event = "transfusion"; event_point1 = "start"; anchor1 = "admission"; anchor_point1 = "start"; relation1 = "equal or after"; event_point2 = "finish"; anchor2 = "discharge"; anchor_point2 = "finish"; relation2 = "equal or before"
4.1.4.3. <i>EMS arrived about 20 minutes after the ingestion</i>		event = "EMS arrived"; event_point = "unspecified"; anchor = "ingestion"; anchor_point = "finish"; relation = "equal"; quantity = "20"*; time_unit = "minute"; direction = "plus"; interval_operator = "jump"; vagueness = "yes"

(continued on next page)

Table 3 (continued)

Statement	Figure	Temporal constraint structure encoding
4.1.5. Fuzzy time		
4.1.5.1. <i>She had appendectomy in the past</i>		event = "appendectomy"; event_point = "unspecified"; anchor = "now"; relation = "before"
4.1.5.2. <i>Diagnosis of astrocytoma was made at that time</i>		event = "diagnosis of astrocytoma"; event_point = "unspecified"; anchor = "previous_reference"; relation = "equal"
4.1.5.3. <i>He subsequently developed a rash.</i>		event = "rash"; event_point = "start"; anchor = "previous_reference"; relation = "after"
4.1.6. Recurring times		
4.1.6.1. <i>Thiamine 100 mg p.o. q.d. for two weeks</i>		event = "Thiamine 100 mg p.o. q.d."; event_point = "start"; anchor = "event"; anchor_point = "finish"; relation = "equal"; quantity = "2"; time_unit = "week"; direction = "minus"; interval_operator = "jump"
4.2. Temporal modifiers and indefinite numbers		
4.2.1. <i>Low WBC lasted for more than four weeks</i>		event = "low WBC"; event_point = "start"; anchor = "event"; anchor_point = "finish"; relation = "before"; quantity = "4"; time_unit = "week"; direction = "minus"; interval_operator = "jump"
4.2.2. <i>He had a fever more than 3 days later</i>		event = "fever"; event_point = "unspecified"; anchor = "narrative_reference"; relation = "after"; quantity = "3"; time_unit = "day"; direction = "plus"; interval_operator = "jump"

Sa: the start of admission; Fd: the finish of discharge; NR: narrative_reference; PR: previous_reference; a line indicates the time over which an event occurs, with a circle on the end indicating the start or the finish is known and without a circle on the end indicating the start or the finish is unspecified. Similar to Fig. 3, there is a difference between the temporal constraint structures and the figures about the length of time. In the temporal constraint structures, we used the exact number of time units as it was stated. However, later on in our postprocessor, we will convert this exact number to a time range to represent the vagueness of the temporal information, which is shown as bars in the figures and marked with * in the structures.

an **anchor** (using ISO 8601 standard to format the time [37]). Importantly, **relation** is typically the temporal preposition located just before the time (e.g. the **relation** for "on 2000-01-01" is encoded as "equal").

We observed features of *time range* and *conjoined time* with date and time expressions. Time range, as demonstrated in example 4.1.1.3., has a beginning and ending point indicating a duration over which an event lasts. This kind of temporal information constrains both the start and the finish of an event using the two endpoints of the time range serving as anchors, respectively. Conjoined time includes those expressions with conjunction (e.g. Example 4.1.1.4.) and disjunction (e.g. "The patient will have a surgery on April 4th or 6th"). A temporal expression with conjunction also gives a range of

time. However, it constrains the start and the finish of an event by locating them at a specific time points within that range. Though disjunction may occur in the context of scheduling (e.g. follow-up plan), we only found one case of disjunction ("Follow up in clinic on 5/2 or 5/9") in 200 discharge summaries.

Several issues regarding granularity and *relations* should also be mentioned. Temporal granularity for historical events was usually *year*. On the other hand, time expressions for most recent events were typically specified using the format *date*. The most frequent prepositions observed within the corpus were *on* (date) and *in* (month/year). Other common prepositions included *until*, *since*, and *by*. Week-day, part-of-day, time-of-day and season information were not commonly observed.

Week-day was often used in the context of a treatment plan; for example, “*Coumadin 2.5 milligrams Monday, Wednesday, Friday*” indicates dosing information for the medication *Coumadin*.

4.1.2. Relative date and time

Examples of relative date and time include “*today*,” “*next year*,” “*two days ago*,” and “*within the last month*.” The reference time for this type of temporal expression varies with the individual report or author. The process of determining the values for the temporal constraint structure requires knowledge of the author’s perspective of time of when the text was composed or whether there was a time reference from the text’s context with a relative date and time to which expressions can refer.

Since hospitalizations usually occur over days, resolving expressions whose granularity is month or year can typically be straightforward. For example, “*last year*” can be resolved relative to discharge date. However, relative date and time with finer granularity, such as “*today*” and “*several minutes ago*,” should probably be interpreted differently according to the relative position or document section of the expression within the discharge summary.

Several example expressions in Table 3 depict relative date and time using relative anchors (e.g. “*narrative_reference*,” indicating the author’s perspective of time when she/he is writing). The temporal constraint structure represents relative time in a manner that attempts to maintain the meaning of original expression as much as possible, so as to make follow-up higher-level processing possible. While beyond the scope of this paper, complex linguistic analysis and domain knowledge are needed to formally resolve these expressions.

Relative date and time temporal expressions can also be nuanced. For instance, the constraint structures of Examples 4.1.2.4. and 4.1.2.5. seem similar. Using world knowledge, we can infer, however, that the event *orthopedic clinic visit* probably took less than a few hours, while the event *decreased appetite* might have persisted over four days, from four days ago to sometime after admission. As above, higher level reasoning systems applying domain knowledge base will be needed to assist in effectively solving this problem.

4.1.3. Duration

Duration is defined as the difference in time between two points with explicit information concerning how long the event lasts. In some cases, only the length of time is known, and neither endpoint can be localized within the time line directly or indirectly using other “reference time” (or previous-reference) (Example 4.1.3.1). In this case, we treated one of the endpoints of the event as an anchor (i.e. the finish point) and constrained the other endpoint (i.e. the

start point). In contrast, other expressions include terms such as “*last*,” “*past*” or “*next*” indicating a direction relative to the *narrative reference*, such as Examples 4.1.3.2. and 4.1.3.3. Example 4.1.3.2 states that hypertension was diagnosed 3 *years ago* from the narrative time point, *admission time*. The length of time of a duration can also be a range (Example 4.1.3.3.), and vague (e.g. *about 3 days*, and *many years*). In clinical reports, durations of treatment are often stated in the format “*seven days of antibiotics*” or “*antibiotics times seven days*.”

4.1.4. Event-dependent temporal expression

Event-dependent temporal expressions use events as anchors. Specifically, endpoint(s) of an event are constrained by relating the endpoint(s) to another event. These expressions can be further divided into two subcategories according to the nature of the reference events using either “key events” or ordinary medical events.

The first subcategory of expressions is based upon “key events,” which are significant events in the clinical environment often used as “time references” by health-care providers. These “key events” include *admission*, *discharge*, *operation*, *transfer*, and *parturition* (Example 4.1.4.1 and 4.1.4.2.). Temporal expressions of this type include “*two days before admission*” and “*on hospital day three*.” With key events, the actual time of the key event is often stated in the report explicitly or can be obtained from other available data sources as a follow-up, higher-level step where the values of these temporal expressions represented within our structure might be fully resolved.

The second subcategory of event-dependent expressions uses ordinary medical events. Though we might not know the exact time of these events, by interpreting these temporal expressions, the relationship between two medical events can be chronologically determined relative to one another. For instance, “*EMS [Emergency Medical Services] arrived about 20 minutes after the ingestion*” (Example 4.1.4.3.), places *EMS arrival* relative to the event *ingestion*.

Temporal conjunctions and prepositions (e.g. “*when*,” “*before*” and “*subsequent to*”) play an important role in linking events relative to one another and provide clues to the presence of this class of temporal expressions. Some causal phrases, such as “*because of*” and “*secondary to*,” also indicate temporal sequences.

4.1.5. Fuzzy time

Temporal expressions are sometimes fuzzy. For example, to determine exact time boundaries for expressions such as “*early morning*” and “*night*” can be difficult. In addition, the anchor for relative date and time may be uncertain. We observed four main categories of fuzzy time expressions.

First, there are a large number of temporal terms referring to the past, present, or future in a general sense (e.g. “*in the past*,” “*now*”). Expressions that refer to the past are mostly used to describe the patient’s medical history; for example, “*she had an appendectomy in the past*.” To represent these temporal expressions, we designate the value of **anchor** as “*now*” with the value of **relation** “*before*” to signify the past (Example 4.1.5.1.).

The second type of fuzzy time expressions is composed of those temporal expressions that do not refer to a specific time. This subclass of expressions would include such statements as: “*his mood was lowest in the morning*” and “*he admits to deep sadness at times*.” The information in these expressions speaks to characteristics of the health problem; the expressions cannot be localized along the time line, and the expressions cannot be used to link events. Therefore, we did not represent these expressions with our temporal constraint structure.

Third, temporal expressions, such as “*at this time*,” “*at that point*,” and “*during this time*,” refer to a time mentioned in previous content. An example of this type of expression is: “*Approximately four years ago, he experienced bifrontal headaches. He visited doctors several times. ... Diagnosis of astrocytoma was made at that time...*” From the context of this example, however, the temporal location for “*at that time*” is ambiguous. For the purposes of representation, **anchor** would be the assigned “*previous_reference*” and **relation** would be assigned “*equal*,” with resolution of these statements perhaps possible with the addition of higher-level linguistic analysis and domain knowledge (Example 4.1.5.2.).

The fourth type of fuzzy temporal expression is the use of temporal adjectives such as “*previous*,” “*subsequent*,” and “*chronic*” and temporal adverbs such as “*subsequently*,” “*afterward*,” and “*lately*.” These temporal adjectives and adverbs indicate whether the events happened before, equal, or after “*now*” (or *previous_reference*) (Example 4.1.5.3.). Some temporal adjectives and adverbs, such as “*gradual*,” “*occasional*,” “*abruptly*,” “*progressively*,” and “*still*,” describes a temporal feature of the medical event itself. We did not intend to represent these expressions using our model, because this descriptive information of the medical event did not specify the temporal ordering of the event.

4.1.6. Recurring times

Some temporal expressions indicate events which occur on multiple occasions, typically at regular intervals. In general, there are several kinds of events which repeat at regular time intervals. The most common example is medication dosing (e.g. “*Thiamine 100 mg p.o. q.d.*”). Theoretically, we can represent these expressions using a series or a nesting of temporal constraint structures. In this example, if we were to consider each medication

administration as a separate event, then the start of the immediately preceding medication dosing is equal to 24 h before the start of next one. We might further assume in this example that the granularity of taking a per oral medication is at the level of seconds; therefore, the finish of the previous dosing is before the start of the next dosing. One difficulty with recurring events is that the number of times that the event actually occurs is not usually known (e.g. skipped doses). Furthermore, the event of medical dosing can also be conceptualized as a treatment plan. Whether the patient has actually taken the medication, however, is unknown. As a result, these events are simply represented as a single interval modeled for the duration, if available. For example, for “*Ampicillin 250 mg q.i.d. for five days*,” we folded the dosing information into the event name as “*Ampicillin 250 mg q.i.d.*,” and modeled the duration “*for five days*” (also see Example 4.1.6.1.).

Some expressions also exhibit irregularly recurring events, including expressions where the distance between each interval is unknown (e.g. “*the patient was somnolent and then combative at times*”) or recurrence is uncertain (e.g. “*he felt guilty almost everyday*”). These expressions are also represented as single intervals.

4.2. Additional temporal modeling issues

The discharge summary corpus also revealed several other features of temporal expressions, including (1) temporal modifiers and indefinite numbers and (2) temporal co-references. While these features do not constitute separate categories of expressions (e.g. a temporal modifier or an indefinite number is a part of a temporal expression, and temporal co-references may relate to two temporal expressions in the same category or across two different categories), they are nevertheless important issues which were formally considered in the temporal constraint structure.

4.2.1. Temporal modifiers and indefinite numbers

Temporal modifiers include: (1) part modifiers (15 of 2022 cases), which usually qualify anchors and are represented by **anchor_modifier** with the values *early*, *mid*, and *late* (e.g. “*in the early eighties*”); (2) quantity modifiers (4 of 2022 cases), which quantify the length of a duration (e.g. “*for more than four weeks*”) and are represented using **relation**; and (3) vague modifiers (60 of 2022 cases) (e.g. “*for approximately six years*”) represented by **vagueness** field with value of “*yes*.” With quantity modifiers, the value of **relation** depends on the value of **direction** in the same temporal constraint structure. For example, if the **direction** is “*minus*,” indicating that the constrained endpoint of an event is “*in the past*” with respect to the anchor, the value of **relation** is *before* for an expression such as “*more than*” (Example 4.2.1), *equal or before* for “*no less than*,” *after* for “*less*

than,” and *equal or after* for “*at most.*” If the value of *direction* is “*plus,*” the value of *relation* should be changed to its opposite (Example 4.2.2).

In addition, temporal expressions can also have unspecified or indefinite numbers associated with their duration. The length of duration sometimes is unspecified (3 of 2022 cases) (e.g. “*he had back pain for years*”). For these types of expressions, the value *quantity* is “*unspecified.*” In addition, durations may also be described with indefinite numbers (64 of 2022 cases) (e.g. “*a few,*” “*a couple,*” “*many*”). Since “*many*” and “*a few*” may represent different amounts, we maintained the original terms as the value of *quantity* in our implementation of the temporal constraint structure.

4.2.2. Temporal co-reference

Temporal expressions can co-refer to one other. In the example, “. . . *on 09/16/2004, the date of admission, the patient had a chief complaint of chest pain. . . .*” *09/16/2004* and *admission* refer to the same time point. In the statement, “*The patient’s medical history was significant for an admission in October 12, 2000 with a chief complaint of bright red blood per rectum. A work up at that time showed sigmoid cancer. . . .*” “*at that time*” refers to “*October 12, 2000.*” For the purposes of our model, each of these references was represented explicitly. Future rules will be needed for later higher-level reasoning systems. In the first case, the more specific expression (the absolute date) might be chosen, while in the second case, linguistic and domain knowledge would be required to accomplish the resolution.

4.3. Temporally-related expressions

Some expressions, including medicine-specific expressions, age, and location, are in part temporally-related. These expressions were not formally modeled in our representation but are discussed here, as they might play a role in future systems which might integrate and include some of these concepts in a higher-level temporal abstraction of the EMR.

4.3.1. Medicine-specific expressions

Medical specific expressions are terms or phrases that contain temporal components but that are not temporal expressions in a formal sense. They can be observed throughout the medical narrative. For example, clinical symptoms such as “*night orthopnea,*” “*morning sickness upon rising*” typically occur during a specific portion of the day. In the statement “*he was assessed with a 24-hour urine,*” “*24-hour urine*” is laboratory test examining the urine produced in a one day period. The statement “*She has fibroid tumors, and her uterus is close to five-month pregnancy size,*” describes the size of a tumor (“*five-month pregnancy*”). Pulse (e.g. *72 beats per minute*) and respiratory rate (e.g. *16 breaths per minute*) are

essentially medical observations, which have regular rates per time unit. These expressions, therefore, have independent meaning, represent specific medical concepts, and do not pin events to a timeline or link events; for these reasons, medical specific expression, while important to consider in medical NLP, were not represented in the temporal constraint structure.

4.3.2. Age

The age of a patient and his/her relatives constitutes a very important piece of temporal information in medicine. It affects not only the diagnosis but also treatment plans and prognosis. As such, clinicians usually declare patient’s age at the beginning of a report. The family history section of a typical history and physical also may contain the ages of a patient’s family members (e.g. “*she has a sister who died of breast cancer at age 36*”). In addition, age can be used to indicate the time in a description (e.g. “*The patient had an appendectomy at age 15*”). Age can also be vague, such as “*since childhood*” or “*during his thirties.*” However, if we know the patient’s birth date, then the calendar time in which the event occurred could be easily calculated. One possible implementation would be to model age as having an anchor that is the patient’s birth date, with a time unit and a quantity indicating the patient’s age.

4.3.3. Location

Location, while not a temporal expression, often indicates temporal change in the medical narrative. In addition, clinicians sometimes use hospital location to organize medical events during a patient’s inpatient stay. The statement, “*he was transferred to the CCU (Coronary care unit) for further management and therapy,*” indicates that events stated before this sentence were temporally before events which occurred in the CCU following this sentence.

4.4. Evaluation of the temporal constraint structure

Overall, 1961 out of 2022 (97%) temporal expressions identified in 100 discharge summaries by a single coder (LZ) were effectively modeled using the temporal constraint structure. Three percent temporal expressions that were not effectively modeled mainly include two categories in Table 1: non-specific time (e.g. his mood was lowest *in the morning*) and recurring time (e.g. he felt guilty almost *everyday*). The subset of 50 expressions evaluated by both coders (LZ, GH) yielded exact match in 90%, partial match with trivial differences in 8% (e.g. the name for an event), partial match with large differences in 2% (e.g. for the statement “*the patient was admitted on 2/3/88 for vomiting of blood,*” both coders encoded admission as an event. However, one coder encoded *vomiting of blood* as another event which

happened before admission, but another coder did not.), and total mismatch in 0%.

5. Discussion

This study provides a methodological representation of temporal expressions for the purposes of aiding the development of future temporal reasoning systems with NLP data in the clinical domain. This study represents one of the few attempts to analyze and model temporal expressions using a large clinical text corpus. Our corpus analysis revealed a wide diversity of temporal expressions in discharge summaries, which we classified into natural categories. These expressions were then characterized by their common features in a concise and expressive way to construct the temporal constraint structure in a grounded approach. The structure was found to provide a comprehensive coverage in encoding varied temporal expressions contained within the corpus.

Several important issues concerning temporal expressions narrative texts were observed in this study. We demonstrated that some temporal information in clinical text have characteristics in common to other domains, such as expressions of date and time. Many temporal expressions from discharge summaries, however, have unique characteristics. By categorizing temporal expressions into different groups and calculating the frequency of each category, our results provide a “big picture” of temporal expressions in discharge summaries, which may benefit future researchers examining temporal information extraction in clinical narratives. As such, certain expressions, which need not be considered for medical narratives (e.g. *century*, *millennia*, and *Before Current Era*), can be eliminated and medically relevant expressions specific to these documents (e.g. “postop day # 4”) can be developed and represented.

Temporal granularity, which is central to temporal reasoning, was found to have interesting characteristics in our corpus of discharge summaries. We found that clinicians use a variety of time granularities but that specific calendar dates were most prevalent. Most durations were based on year or day. Our analysis also revealed a set of key events which could often serve as temporal anchors, including *admission*, *discharge*, and *operation*. While further investigations would be needed, these observations and other work [38] suggest that in a particular clinical domain, such as that surrounding a specific disease (e.g. *acute myocardial infarction*) or a certain therapy management (e.g. *deep venous thrombosis prophylaxis*), a set of domain-specific key events can be identified to assist in temporal inference.

One of the central motivations of our structured data approach was to provide a means by which temporal reasoning systems can integrate further knowledge, pro-

duce important intermediate downstream constructs, and ultimately derive general temporal implication. Furthermore, having a structured representation means that these formatted temporal expressions can be stored in a clinical data warehouse and more easily used for diverse medical informatics applications, such as information retrieval, data summarization, temporal abstraction, and other medical research.

Placing this work into an overall context, temporal representation and reasoning using natural language data have been associated with many fields, including natural language understanding and processing, artificial intelligence, and medical science. Perhaps the two best known attempts at representing temporal expressions are the formalisms proposed by the TIDES group [39], which developed a standard for the annotation of temporal expressions, and the TimeML group [40,41], which developed a language for representing events and temporal expressions in natural language. While these formalisms represent important contributions to the general field of temporal reasoning, their work cannot be directly compared, primarily because the text corpus used by these groups was news articles.

As previously established, natural language in medicine is more restricted and well-defined than the general domain [42], as it has been demonstrated to have a specific vocabulary, set of semantic relations, and, in some cases, syntax. Although the comprehensive treatment of temporal issues is essential to improving the performance of medical NLP, implemented systems have used instant-based, absolute-time frameworks [1,2]. Our proposed representation has the potential to enable automated annotation and extraction of temporal information by NLP systems.

One of the simple and powerful aspects of our proposed representation model is that it maintains the original meaning of most expressions using a set of consistent fields. This representation therefore maximizes the expressive capability of the expressions placed into the structure. The structure also successfully connects temporal information with medical events by modeling the time over which a medical event occurs as in interval in which the ending time point(s) constrain each medical event.

As a next step, the temporal constraint structure could be integrated with current medical NLP systems. At our institution, for example, MedLEE [15,16,18] has been applied to process a wide range of medical texts. It uses a frame-based representation and employs a vocabulary and a grammar to extract information for narrative text. The statement “his cough started two weeks before admission” is encoded by MedLEE using a limited temporal schema previously. The temporal constraint structure, allows for more sophisticated temporal encoding which maintains the original meaning of

the expression. Using the same example, the structured output would be:

```

problem: cough
temporal constraint:  event_point = "start"
                      relation = "before"
                      quantity = "2"
                      time_unit = "week"
                      direction = "minus"
                      interval_operator = "jump"
                      anchor = "admission"
                      anchor_point = "start"

```

We have implemented this as a pre-processing step so that the design of MedLEE would not need to be modified, as the temporal constraint structure is added as a modifier to the primary information in the top level frames.

The addition of several areas of knowledge and reasoning, which will be required to ultimately accomplish temporal resolution using constraint-based approaches such as the STP, is facilitated with this representation structure that provides temporal information in an accessible, coded format. As was illustrated in our treatment of the discharge summary corpus, temporal information is often vague. Some temporal expressions have vagueness qualifiers such as “*about*” or indefinite numbers such as “*many*.” Previously, we modeled vagueness by widening the limits of the constraints based on our experience [24]. We believe that rules should be made depending upon textual context, the specific study domain, and with increasing research experience. As such, vagueness is not directly represented in the temporal constraint structure, and we attempt to maintain the original meaning of the temporal expressions, so that higher-level temporal reasoning systems can have maximal flexibility for dealing with these expressions.

Similarly, while the majority of constraint structures included in our results was derived from explicit temporal assertions, assertions in natural language are often implicit. Some constraints are based on domain knowledge and assumptions. For example, *death* is instantaneous, so the finish of *death* is equal to the start. To obtain this implicit information, linguistic and domain knowledge must be utilized based on the source of the data and the specific medical domain. In addition, further processing and addition of higher-level domain knowledge need to be added to resolves expressions such as “*now*,” “*today*,” or other expressions referred in our representation as narrative references.

The medical events that we modeled were historical events. Discharge summaries contain some hypothetical events, such as therapeutic plans, follow-up plans, and prognosis. Instead of modeling these events separately, we represent *the statement of these hypothetical events*. While beyond the scope of this paper, NLP techniques, such as temporal discourse analysis in conjunction with

medical domain knowledge, may be required to resolve some of these issues.

We have previously demonstrated that temporal expressions in discharge summaries can be modeled as a simple temporal problem [24]. Temporal assertions encoded in the proposed temporal constraint structure can easily be used with the addition of reasoning formalisms based on constraint-based techniques, especially STPs. An STP involves a set of variables, X_1, \dots, X_n . Each variable represents a time point, where a time point may be a start or end time of some events. A constraint bounds the distance between two time points and could be denoted as $a \leq X_i - X_j \leq b$, where X_i and X_j are two time points, interval $[a, b]$ represents the constraint, and the inequalities can be strict and non-strict, $<$ and \leq , respectively. A special time point, X_0 is “the origin.” Times are relative to X_0 ; thus a point X_i in absolute time can be anchored with the constraint: $d \leq X_i - X_0 \leq d$, where d is the duration from X_0 to X_i . An STP can be represented by a distance graph, $G_d = (V, E_d)$, in which vertices represent variables and each edge is labeled by an interval representing the constraint in inequalities form as shown in Fig. 1. Solving an STP involves finding a solution of linear inequalities on the variables. A shortest path algorithm is applied to solve the linear inequalities. The time complexity of solving STPs is $O(n^3)$.

In our current representation, the fields *event_point* and *anchor* are variables X_i and X_j , representing time points. The constraint structure specifies a set of intervals (e.g. (14, ∞) means greater than 14 time units) and constrains time points. As such, the constraint structure for the statement “*his cough started at least two weeks before admission*” shown before would be translated to a constraint represented by an interval: $14 \text{ days} \leq X_{\text{start_of_admission}} - X_{\text{start_of_cough}} \leq \infty$. The set of points and constraints from a set of expressions can then be represented in a network as shown in Fig. 1. By applying so-called all-pairs-shortest-paths algorithm, the consistency of the network is checked and more temporal information is produced [24]. The results of our current analysis revealed that the great majority of events in discharge summaries are historical and without temporal disjunctions, so the STP would be sufficient to represent temporal assertions in discharge summaries.

It is important to also note that the temporal constraint structure can support other formalisms beyond the STP. Temporal disjunctions can be represented using this structure by simply noting “or” among the related temporal structures. Therefore, the temporal constraint structure could support the general TCSP, and so could be used even if it is necessary to provide a more expressive way of modeling time than we have described here. Similarly, constraint networks in Vilain and Kautz’s point algebra [32] are a special case of a TCSP, lacking metric information, which could be supported with the constraint structure by ignoring

durations. Moreover, Allen's interval algebra represents the start and end times of an event with a pair of time points (X_i and X_j). The allowed relations between pairs of variables are taken from the 13 interval relations (before, meets, overlaps, during, starts, finishes, and their inverses and equal). As such, our temporal constraint structure could be implemented to model a constraint between two Allen intervals (e.g. A equal B), where $A = [X_m, X_n]$ and $B = [X_p, X_q]$ using four variables (e.g. (X_m equal X_p) and (X_n equal X_q)) [34]. In other words, the representation we have introduced here can capture the temporal information in all commonly used temporal representations.

A limitation of this study is its exclusive use of discharge summaries. While this study does not explicitly examine other types of electronically available clinical texts such as radiology reports, pathology reports, and operative notes, it has been our experience that these reports are relatively short and contain much simpler temporal information compared with discharge summaries. Most temporal expressions in these documents are absolute date or time expressions, and it is our belief that this model could be easily applied to these other reports. Another limitation is that both coders for our evaluation are also co-developers, which might affect the assessment of the reliability of coding. Therefore, the agreement in this study might be higher than using other coders.

6. Conclusion

The proposed temporal constraint structure for representing medical temporal expressions was able to effectively encode most temporal expressions contained within discharge summaries. The structure highly preserves the original meaning of temporal information in a complete and concise manner. Placing data within the structure provides a foundational representation upon which further reasoning, including the addition of domain knowledge and other post-processing to implement a STP, can be accomplished.

Acknowledgments

This work was funded by National Library of Medicine grants R01 LM06910 "Discovering and applying knowledge in clinical databases"; R01 LM07659 "Capturing and Linking Genomic and Clinical Information;" and R01 LM07268 "Using Narrative Data to Enrich the Online Medical Record."

References

[1] Augusto JC. Temporal reasoning for decision support in medicine. *Artif Intell Med* 2005;33(1):1–24.

[2] Combi C, Shahar Y. Temporal reasoning and temporal data maintenance in medicine: Issues and challenges. *Computers in Biology and Medicine* 1997;27(5):353–68.

[3] Kindler H, Densow D, Fliedner TM. A pragmatic implementation of medical temporal reasoning for clinical medicine. *Comput Biol Med* 1998;28(2):105–20.

[4] Deshpande AM, Brandt C, Nadkarni PM. Temporal query of attribute-value patient data: utilizing the constraints of clinical studies. *Int J Med Inform* 2003;70:59–77.

[5] Nigrin DJ, Kohane IS. Temporal expressiveness in querying a time-stamp-based clinical database. *J Am Med Inform Assoc* 2000;7(2):152–63.

[6] Shahar Y, Musen MA. Knowledge-based temporal abstraction in clinical domains. *Artif Intell Med* 1996;8(3):267–98.

[7] Shahar Y. A framework for knowledge-based temporal abstraction. *Artif Intell* 1997;90(1-2):79–133.

[8] Kahn M, Fagan L, Sheiner L. Combining physiologic models and symbolic methods to interpret time-varying patient data. *Methods Inform Med* 1991;30(3):167–78.

[9] Das A, Musen M. A foundational model of time for heterogeneous clinical databases. *Proc AMIA Annu Fall Symp* 1997.

[10] Combi C, Pincirolu F, Pozzi G. Managing different time granularities of clinical information by an interval-based temporal data model. *Methods Inform Med* 1995;34(5):458–74.

[11] Combi C, Pincirolu F, Pozzi G. Managing time granularity of narrative clinical information: the temporal data model TIME-NESIS. In *Proceedings of the 3rd workshop on temporal representation and reasoning (TIME'96)*. p. 88.

[12] O'Connor M, Tu S, Musen M. Representation of temporal indeterminacy in clinical databases. *Proc AMIA Symp* 2000;1:615–9.

[13] Campbell KE, Das AK, Musen MA. A logical foundation for representation of clinical data. *J Am Med Inform Assoc* 1994;1(3):218–32.

[14] Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest X-ray reports. *J Am Med Inform Assoc* 2000;7(6):593–604.

[15] Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. *J Am Med Inform Assoc* 1994;1(2):161–74.

[16] Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking clinical data from narrative reports: a study of natural language processing. *Ann Intern Med* 1995;122(9):681–8.

[17] Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. *J Am Med Inform Assoc* 2005;M1794.

[18] Friedman C. A broad-coverage natural language processing system. *Proc AMIA Symp* 2000:270–4.

[19] Sager N, Lyman M, Nhan N, Tick L. Medical language processing: applications to patient data representation and automatic encoding. *Methods Inform Med* 1995;34(1-2):140–6.

[20] Wilcox A, Hripcsak G. The role of domain knowledge in automating medical text report classification. *J Am Med Inform Assoc* 2003;10:330–8.

[21] Cao H, Stetson P, Hripcsak G. Assessing explicit error reporting in the narrative electronic medical record using keyword searching. *J Biomed Inform* 2003;36:99–105.

[22] Mendonca EA, Hass J, Shagina L, Larson E, Friedman C. Extracting information on pneumonia in infants using natural language processing of radiology reports. *J Biomed Inform* 2005;38(4):314–21.

[23] Dolin RH. Modeling the temporal complexities of symptoms. *J Am Med Inform Assoc* 1995;2(5):323–31.

[24] Hripcsak G, Zhou L, Parsons S, Das AK, Johnson SB. Modeling electronic discharge summaries as a simple temporal constraint satisfaction problem. *J Am Med Inform Assoc* 2005;12(1):55–63.

- [25] Jonsson P, Drakengren T, Backstrom C. Temporal information in medical narrative. In: Sager N, Friedman C, Lyman MS, editors. Medical language processing computer management of narrative data. Reading, MA: Addison-Wesley Pub Co.; 1987. p. 175–94.
- [26] Zhou L, Friedman C, Parsons S, Hripcsak G. System architecture for temporal information extraction representation and reason in clinical narrative reports. Proc AMIA Symp 2005. [accepted].
- [27] Ceusters W, Buekens F, DeMoor G, Bernauer J, DeKeyser L, Surján G. TSMI: a CEN/TC251 standard for time specific problems in healthcare informatics and telematics. 1997;46(2):87.
- [28] Allen JF. Time and time again: the many ways to represent time. Int J Intell Syst 1991;6(4):341–55.
- [29] Pani AK, Bhattacharjee GP. Temporal representation and reasoning in artificial intelligence: a review. Math Comput Model 2001;34(1–2):55–80.
- [30] Meiri I. Combining qualitative and quantitative constraints in temporal reasoning. Artif Intell 1996;87(1–2):343–85.
- [31] Dechter R, Meiri I, Pearl J. Temporal constraint networks. Artif Intell 1991;49(1–3):61–95.
- [32] Vilain M, Kautz H. Constraint propagation algorithms for temporal reasoning. In Proceedings AAAI-86, Philadelphia, PA; 1986. p. 377–82.
- [33] Dean T, McDermott D. Temporal data base management. Artif Intell 1987;32:1–55.
- [34] Allen J. Maintaining knowledge about temporal intervals. Commun ACM 1983;26:832–43.
- [35] Xu L, Choueiry B. A new efficient algorithm for solving the simple temporal problem. In Proceedings of 10th international symposium on temporal representation and reasoning and fourth international conference on temporal logic, July 08–10, 2003, Cairns, Queensland, Australia; 2003. p. P212.
- [36] Lisa Ferro LG, Inderjeet Mani, Beth Sundheim, George Wilson. TIDES—2003 standard for the annotation of temporal expressions. In Proceedings of the MITRE 2003.
- [37] ISO8601:2000(E) Data elements and interchange formats—information interchange—representation of dates and times. 2nd ed.; 2000-12-15.
- [38] Obermeier K. Temporal inference in medical texts. In Proceedings of 23 annual meeting of the Association for Computational Linguistics, Chicago; 1985 July.
- [39] Ferro L, Gerber L, Mani I, Sundheim B, Wilson G. TIDES 2003 standard for the annotation of temporal expressions, September 2003;2003.
- [40] Sauri R, Littman J, Knippen B, Gaizauskas R, Setzer A, Pustejovsky J. TimeML annotation guidelines; September 2004.
- [41] Hobbs J, Pustejovsky J. Annotating and reasoning about time and events. In Proceedings of AAAI spring symposium on logical formalizations of commonsense reasoning, Stanford, CA; March 2003.
- [42] Friedman C, Kra P, Rzhetsky A. Two biomedical sublanguages: a description based on the theories of Zellig Harris. J Biomed Inform 2002;35(4):222–35.