Temporal Cohort Logic

Guo-Qiang Zhang[1,2,3], Xiaojin Li[1,3], Yan Huang[1,3], Licong Cui[2,3]
[1]McGovern Medical School, [2]School of Biomedical Informatics, [3]Texas Institute for Restorative Neurotechnologies
The University of Texas Health Science Center at Houston, Houston, Texas 77030, USA

Abstract

We introduce a new logic, called Temporal Cohort Logic (TCL), for cohort specification and discovery in clinical and population health research. TCL is created to fill a conceptual gap in formalizing temporal reasoning in biomedicine, in a similar role that temporal logics play for computer science and its applications. We provide formal syntax and semantics for TCL and illustrate the various logical constructs using examples related to human health. Relationships and distinctions with existing temporal logical frameworks are discussed. Applications in electronic health record (EHR) and in neurophysiological data resource are provided. Our approach differs from existing temporal logics, in that we explicitly capture Allen's interval algebra as modal operators in a language of temporal logic (rather than addressing it in the semantic structure). This has two major implications. First, it provides a formal logical framework for reasoning about time in biomedicine, allowing general (i.e., higher-levels of abstraction) investigation into the properties of this approach (such as proof systems, completeness, expressiveness, and decidability) independent of a specific query language or a database system. Second, it puts our approach in the context of logical developments in computer science, allowing potential translation of existing results into the setting of TCL and its variants or subsystems so as to illuminate opportunities and computational challenges involved in temporal reasoning for biomedicine.

1 Introduction

Discovery in life science today is being enabled through computational- and data-intensive research that exploits the enormous amounts of available biomedical and health data. Real-world data (RWD) such as electronic health record (EHR) offer tremendous opportunities for human health research. Such opportunities include traditional retrospective analyses for identifying risk profiles, revealing health disparities, and understanding long-term health implications. Recent advances in data science and machine learning open further opportunities in areas such as clinical decision support, outcome predication, drug repurposing, and poly-pharmacy [1, 2].

The Office of National Coordinator for Health Information Technology reports that, between 2008 and 2015, the adoption of basic EHR technology in the United States rose from about 10 to over 80 percent [3]. To leverage RWD for research, academic medical centers created and maintain Enterprise Data Warehouses (EDWs) or integrated clinical data repositories (IDRs) combining multiple facets of data generated from patient care into a "single source of truth." Cohort discovery is the process involved in identifying and extracting data on subgroups of patients from an IDR that are suited for generating real-world evidence (RWE) from RWD to support study objectives. This process is often supported by query engines with a graphical user interface (e.g., [4, 5] designed to ease the effort involved in data exploration without requiring the knowledge about how the backend data are stored and managed.

Temporal query, an important aspect of cohort discovery, has not been the traditional focus of query interfaces [6]. Temporal queries on large EHR-derived datasets presents an emerging big data challenge, since they can be conceptually complex and computationally expensive. For example, queries such as "finding all patients who developed neurologic complications after extracorporeal membrane oxygenation for Covid-19," or "finding all patients who did not have any cardiovascular condition before positive Covid-19 diagnosis" involve complex temporal reasoning for them to be faithfully translated into database queries.

There is, however, *a lack of formalized temporal logic frameworks that can readily serve as a logical foundation for temporal reasoning involve RWD for clinical and population health research.* The wealth of achievements in logics for computer science since the 70's have focused on the specification and verification of properties of computational systems and processes. For the most part, these achievements are not directly applicable to the emerging area of applying RWD for clinical and population health research.

The goal of this paper is to introduce Temporal Cohort Logic (TCL) as a starting point to fill this gap, with demonstrative application use cases. Our overall conceptual setup involves the treatment of

- Clinical terminologies and codes for documenting patient care as atomic propositional variables for TCL syntax;
- Patient medical histories as semantic structures;

- Operators in Allen's interval algebra (e.g. "before," "during" [7]) as temporal modalities in TCL;
- Our model of time correspond to "discrete linear time," although alternative models are possible if needed; and
- The computational challenge shifts from "model-checking" an individual, complex model to model-checking massive amounts of simple models (one may call this "massive model-checking").

Figure 1 illustrates how a patient's clinical history (above) can be lined up using natural numbers (below) in the abstract sense. Here, the granularity of time is days, although more refined time scale (e.g. seconds) can be treated in similar ways as application context warrants, without losing generality.

The main contributions of this paper are: 1. we introduce Temporal Cohort Logic to fill a conceptual gap in formalizing temporal reasoning in biomedicine; 2. we provide a formal treatment of the syntax and semantics for TCL, which incorporates Allen's temporal operators as modalities in the logic; 3. we instantiate TCL in the application temporal query interfaces involving EHR and neurophysiological data; and 4. we discuss the implications of TCL in relationship to recent data science challenges and advances.

These contributions have two major implications. First, TCL allows general (i.e., higher-levels of abstraction) investigation into the properties of this framework independent of a specific query language or a database system. Second, it puts our approach in the context of logical developments in computer science (from the 70's to date), allowing the translation of existing results into the setting of TCL and its variants or subsystems so as to illuminate possible further advances in temporal reasoning for biomedicine.

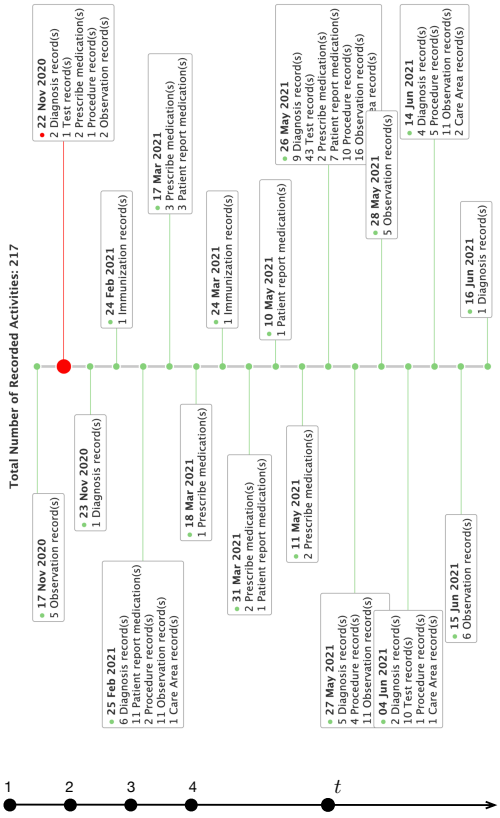

Figure 1: Above: timeline display from a patient's clinical record in an EHR. The number of records on each day is displayed. Below: projection of such a record in a discrete, linear-time axis.

2 Background

2.1 Temporal Logic

Temporal logic is a type of modal logic [8, 9] which has been extensively developed and applied in computer science for specifying and verifying properties about sequential and concurrent programs and systems [10, 11, 12, 13] . It has also been used as a formalism for clarifying philosophical issues, for investigating the semantics of temporal expressions in natural languages, and for capturing temporal knowledge in artificial intelligence.

The basic modality of temporal logic include time-related operators such "sometimes" ($\Diamond$ in notation) and "always" ($\Box$ in notation). Many other types of modalities have been used for formal representation and reasoning about time and temporal information within a logical framework, such as "next time" and "until." First-order and higher-order extensions further increases the expressive power.

The precise meaning of formulas in temporal logics are reflected in their formal semantics, usually interpreted in mathematical structures such as Kripke frames, labeled transition systems, or automata [9, 11]. Nodes (or states) in such structures represent basic temporal units (as a time point or time interval) and temporal changes are captured as relations among the states. Different modal operators in syntax, coupled with different classes of semantic structures allow a rich variety of temporal logics that include Interval temporal logic (ITL [14]), Linear temporal logic (LTL [15]), Hennessy-Milner logic (HML [17]), and Timed propositional temporal logic (TPTL [18]). An important question in temporal logic, which has had industrial applications in hardware design, is model checking [12, 13]. Model-checking is concerned with efficient algorithms verifying that a certain hardware design (represented as a semantic structure $\mathcal{M}$) meets intended properties (represented as a temporal logic formula φ), expressed as $\mathcal{M} \models \varphi$.

Figure 2: Illustration of Allen's classical temporal relations between intervals (pictured in black and in blue) on the real line.

2.2 Allen's Interval Algebra and Halpern-Shoham Logic

Allen's Interval Algebra (Figure 2 [7]) is motivated with the need for AI systems to model space and time in a qualitative, human-like, manner. The basic unit of the algebra is intervals on the real line, which can represent the duration of events, tasks, or actions over time. This algebra formalizes relations such as precedes (i.e., before) and overlaps to encode the possible configurations between those intervals. Allen's Interval Algebra has been used primarily as a qualitative constraint languages, with applications that involve planning and scheduling, natural language processing, temporal databases, and multimedia databases.

Halpern-Shoham logic (HS [19]) is perhaps the only existing, but well-known, logical framework that incorporates Allen's Interval Algebra as a formal component of the logic. HS contains modal operators representing Allen's binary relations between intervals that include: begins (B), during (D), ends (E), overlaps (O), adjacent to (A), later than (L), and their converses. For example, $<B>\varphi$ reads that "there is an interval beginning the current interval, in which φ holds." Therefore, the basic building block of reasoning in HS logic is an interval.

The Temporal Cohort Logic introduced in this paper, although also capturing Allen's notion of interval relations, is distinct from HS in several major ways. *First and perhaps most important,* we do not enforce "intervals" as the basic unit of thought. Instead, we use discrete linear time, the most commonly used time model, as the basic unit of our logical framework. *Second,* motivated by EHR and RWD applications, time-intervals are not the "first-class citizen" of our semantic model. Instead, it is at the derived, monochromatic substructure level, that we describe Allen's interval relations. *Third,* the primary question about TCL is not "satisfiability" or "model-checking." Instead, motivated from population health and biomedical applications, we are interested in specifying and extracting cohorts (hence the name Cohort Logic) from RWD that can serve study objectives. Therefore, we treat each study subject (or data record) as a semantic structure σ, and globally check if σ satisfies φ (i.e., $\sigma \models \varphi$?), instead of focusing on a specific time point t if a study subject's medical history satisfies the properties specified in a TCL formula (i.e, $(\sigma, t) \models \varphi$?). The collection of all study subjects that satisfies φ forms a cohort set $[\![\varphi]\!]$, defined as $\{\sigma \mid \sigma \models \varphi\}$.

2.3 EHR Data, Their Data Models and Query Interfaces

Electronic health record represents a valuable data source that can facilitate clinical and population research. Typical EHR data contain five types of core healthcare encounter information about a patient: *demographics, diagnosis, medication, lab test, and procedure.* It provides a view of the healthcare encounter activities of an individual to facilitate the delivery of care by physicians and other professionals.

Before such data can be effectively used for research, an Extract, Transform and Load (ETL) process is often involved in transforming and integrating data from multiple information systems in a common target data model, exemplified by i2B2, PCORNet, or OMOP. There have been systems developed for querying longitudinal clinical data sources using such data models, such as TriNetX, Atlas for cohort generation with standardized observational data converted to the OMOP Common Data Model [20], Leaf [21], and GENE2D for genetic disorders data [22]. Ontology-driven interfaces include VISAGE [23], MEDCIS [24], DataSphere [5], x-search [25], and ACE [26].

However, support for temporal query remains severely under developed in this context. A key challenge is a gap in conceptual framework: almost all existing approaches involve a direct translation of front-end query specification to database queries. A higher-level, logical system is needed to systematically and uniformly mediate this translation while providing a common formal language supporting temporal expression and reasoning.

The Temporal Cohort Logic introduced in this paper serves to fill this gap. It uses the core healthcare encounter information mentioned above as the basic building block: *diagnosis, medication, lab test, and procedure* are treated in TCL as sets of atomic propositional variables. Demographics can also be treated as atomic propositional variables (with age derived from date of birth), with the unique property that demographic attributes of a patient are assumed to not change overtime. By modeling of each patient's EHR history as a semantic structure for TCL, we can determine, at each given time, if a diagnosis, medication, lab test, or procedure holds true or not according to their role as propositional variables. This treatment is further facilitated by the fact that controlled vocabularies have been used for EHR. For example, ICD-10 has been used for diagnosis, National Drug Code (NDC) has been used for medication, CPT code has been used for procedure, and LOINC has been used for lab test. Such codes have been further mapped into more comprehensive terminology systems such as SNOMED CT, as well as in larger ontological systems such as Unified Medical Language System, making it possible to formulate TCL propositions at different levels of granularity and incorporating background knowledge to facilitate logical inference.

3 Methods

Our overall formulation of Temporal Cohort Logic follows the Tarskian style. It has three main components: (1) *An abstract syntax* that defines the well-formed logical formulas φ; (2) *A class of mathematical structures* describing the semantic space (or frame, or structure $\mathcal{M}$) that the logical formula can be evaluated for its truth status; and (3) *A satisfaction relation* which specifies exactly when a state in a semantic structure holds true $(\mathcal{M}, \sigma) \models \varphi$.

3.1 Syntax

Example temporal formulas that TCL is designed to capture include: *"Diagnoses of COVID-19 by PCR test before COVID-19 vaccination," "ICU stay during hospitalization,"* and *"Intracerebral hemorrhage started by heart attack."*

Neurophysiological examples include

Administration of Ativan, Lorazepam, or Diazepam after end of seizure,
Observation of Sign of Four during a seizure, and
Generalized tonic phase followed by a clonic phase during a seizure, and EEG suppression after the end of seizure.

Formally, TCL formulas are defined using Backus-Naur Form (BNF), a concise description of a context-free grammar:

$$\varphi, \psi ::= p \mid \neg\varphi \mid \varphi \wedge \psi \mid \varphi \vee \psi \mid \varphi X \psi$$

Atomic formulas p are drawn from a predefined finite set Var of atomic propositional variables ($p \in \mathsf{Var}$). Binary temporal operators X are drawn from the collection $\{\mathsf{A}, \mathsf{L}, \mathsf{B}, \mathsf{E}, \mathsf{D}, \mathsf{O}\}$, with respective intended denotations as: A for "meets," L for "before," B for "started-by," E for "finished-by," D for "contains," and O for "overlaps." To both relate to and differentiate from HS logic, we use the same temporal modalities syntactically. Classical boolean logic operators are included in the syntax: $\neg$ for "not" or negation; $\wedge$ for "and" or conjunction; and $\vee$ for "or", or disjunction.

For example, "Intracerebral hemorrhage started by heart attack" is expressed by `I61` B `I219`, where `I61` is the ICD-10 code "nontraumatic intracerebral hemorrhage," and `I219` is the ICD-10 code for "acute myocardial infarction."

A few remarks are in order. First, the BNF notation implies that arbitrary numbers of nesting of logical and temporal modalities are allowed. For example, $(p \wedge q)\mathsf{E}((q\mathsf{A}r) \vee (p\mathsf{D}\neg r))$ is a well-formed formula for TCL, although what it intends to capture is a different matter. Second, while our temporal modalities are binary (in the form $\varphi X \psi$), HS logic uses unary modalities in the form $X\psi$, intended for an existing interval in the semantic structure to be evaluated with respect to an interval satisfying ψ. Third, binary temporal modalities have been previous studied, such as $\varphi\mathsf{U}\psi$ for φ "until" ψ, although not for Allen's operators.

3.2 Formal Semantics

To capture the formal semantics of TCL, we describe the mathematical structures in which TCL formulas can be interpreted or evaluated. Our notion of time is *discrete linear time*, defined using order-theoretic properties. An ordered set $(D, \leq)$ is linear if it is a total order, in which for any two elements x, y, either $x \leq y$, or $y \leq x$. It is discrete in the sense that each element has a successor (called next-time). We consider $(D, \leq)$ to be well-founded, in the sense that there is a minimal element in D (without predecessor), although this is only a matter of convenience.

Without loss of generality, we define a semantic structure to be a mapping $\sigma : \mathbb{N} \to 2^{\mathsf{Var}}$ (called an *assignment*) where $\mathbb{N} = \{1, 2, 3, \cdots, t, \cdots\}$ is the set of natural numbers (a prototypical well-founded discrete, total order), and 2^{Var} is the power set of atomic propositions. Intuitively, each σ corresponds to a sequence of finite subsets (empty set allowed) over Var: $V_1, V_2, \cdots, V_t, \cdots$, which represents a patient's medical history or other study item's recorded properties over time. We deliberately avoided tying time unit to elements of $\mathbb{N}$, although for EHR we can use "days" as the unit, and for EEG signals we can think of it as "seconds." Different application contexts may suggest different types of temporal units.

With these preparation, the meaning of TCL formulas can be specified as follows with respect to σ and $t \in \mathbb{N}$. For the component of propositional logic, we define, as usual (the "if" part suffices in a definition):

$$
\begin{array}{l}
(\sigma, t) \models p \textit{ if and only if } p \in \sigma(t); \\
(\sigma, t) \models \neg\varphi \textit{ if and only if } (\sigma, t) \not\models \varphi; \\
(\sigma, t) \models \varphi \wedge \psi \textit{ if and only if } (\sigma, t) \models \varphi \textit{ and } (\sigma, t) \models \psi; \textit{ and} \\
(\sigma, t) \models \varphi \vee \psi \textit{ if and only if } (\sigma, t) \models \varphi \textit{ or } (\sigma, t) \models \psi.
\end{array}
$$

As usual, one can encode logical implication $\rightarrow$ as $\varphi \rightarrow \psi := \neg\varphi \vee \psi$. For temporal operators, we specify their semantics by establishing the relation $\models$ between (σ, s) and TCL formulas, inductively, as follows (see Figure 3 for graphical illustration).

Meets: $(\sigma, s) \models \varphi \,\mathsf{A}\, \psi$ if there exist $t_0 > s \in \mathbb{N}$ such that $(\sigma, t_0) \models \varphi$, $(\sigma, t_0 + 1) \models \psi$, and

$$
\begin{array}{l}
\textit{for any } t > s \in \mathbb{N}, \textit{ if } (\sigma, t) \models \varphi \textit{ then } t \leq t_0; \textit{ and} \\
\textit{for any } t > s \in \mathbb{N}, \textit{ if } (\sigma, t) \models \psi \textit{ then } t \geq t_0 + 1.
\end{array}
$$

Before: $(\sigma, s) \models \varphi \,\mathsf{L}\, \psi$ if there exist $t_0, t_1 \in \mathbb{N}$ with $s < t_0 < t_1$ and $(\sigma, t_0) \models \varphi$, $(\sigma, t_1) \models \psi$, such that

$$
\begin{array}{l}
\textit{for any } t > s \in \mathbb{N}, \textit{ if } (\sigma, t) \models \varphi \textit{ then } t \leq t_0; \textit{ and} \\
\textit{for any } t > s \in \mathbb{N}, \textit{ if } (\sigma, t) \models \psi \textit{ then } t \geq t_1.
\end{array}
$$

Started-by: $(\sigma, s) \models \varphi \,\mathsf{B}\, \psi$ if there exist $t_1 > s \in \mathbb{N}$ with $(\sigma, t_1) \models \varphi \wedge \psi$ such that

$$
\begin{array}{l}
\textit{for any } t > s \in \mathbb{N}, (\sigma, t) \models \psi \rightarrow \varphi; \textit{ and} \\
\textit{for any } t < t_1 \in \mathbb{N}, (\sigma, t) \models \varphi \rightarrow \psi.
\end{array}
$$

Finished-by: $(\sigma, s) \models \varphi \,\mathsf{E}\, \psi$ if there exist $t_0 > s \in \mathbb{N}$ with $(\sigma, t_0) \models \varphi \wedge \psi$, such that

$$
\begin{array}{l}
\textit{for any } t > s \in \mathbb{N}, (\sigma, t) \models \psi \rightarrow \varphi; \textit{ and} \\
\textit{for any } t > t_0 \in \mathbb{N}, (\sigma, t) \models \varphi \rightarrow \psi.
\end{array}
$$

Contains: $(\sigma, s) \models \varphi \,\mathsf{D}\, \psi$ if there exist $t_0 > s \in \mathbb{N}$ with $(\sigma, t_0) \models \psi$, such that for any $t > s \in \mathbb{N}$, $(\sigma, t) \models \psi \rightarrow \varphi$.

Overlaps: $(\sigma, s) \models \varphi \,\mathsf{O}\, \psi$ if there exist $t_0 > s \in \mathbb{N}$ with $(\sigma, t_0) \models \varphi \wedge \psi$ such that

$$
\begin{array}{l}
\textit{for any } t > t_0 \in \mathbb{N}, (\sigma, t) \models \varphi \rightarrow \psi; \textit{ and} \\
\textit{for any } t < t_0 \in \mathbb{N}, (\sigma, t) \models \psi \rightarrow \varphi.
\end{array}
$$

The intended semantics of these TCL modal operators can be understood as Allen's interval relations on the monochromatic subgraphs induced by the respective TCL formulas (Figure 3). A subset $X \subseteq \mathbb{N}$ is said to be φ-monochromatic with respect to σ if $(\sigma, t) \models \varphi$ for all $t \in X$. A φ-induced monochromatic set, or φ-set in short and $[\![\varphi]\!]_\sigma$ in notation, is defined as the set $\{t \in \mathbb{N} \mid (\sigma, t) \models \varphi\}$.

For example, the satisfaction definition for "started-by," $(\sigma, s) \models \varphi \,\mathsf{B}\, \psi$, can be interpreted as: "φ-set is started by ψ-set in the cofinite segment $[s, \infty)$." The semantic definition above ensures that within the cofinite segment $[s, \infty)$, there exists an interval $[t_0, t_1]$ such that t_0 and t_1 are both in $[\![\varphi]\!]_\sigma$ and $[\![\psi]\!]_\sigma$; $[t_0, t_1] \cap [\![\varphi]\!]_\sigma = [t_0, t_1] \cap [\![\psi]\!]_\sigma$; $[\![\psi]\!]_\sigma \subseteq [\![\varphi]\!]_\sigma$, and $[s, t_0) \cap [\![\varphi]\!]_\sigma$ as well as $[s, t_0) \cap [\![\psi]\!]_\sigma$ are both empty. Similar induced monochromatic set interpretations are possible for other temporal operators.

The key distinction from the interpretation of Allen's interval operator or HS logic is that we do not require φ-set to be consecutive (or closed), in the sense that if t falls in between two members in $[\![\varphi]\!]_\sigma$, it is not necessarily the case that t is also a member in $[\![\varphi]\!]_\sigma$. Intuitively, $[\![\varphi]\!]_\sigma$ can be "porous," to reflect the general situation of non-consecutive events of the same kind taking place, sometimes sporadically, overtime.

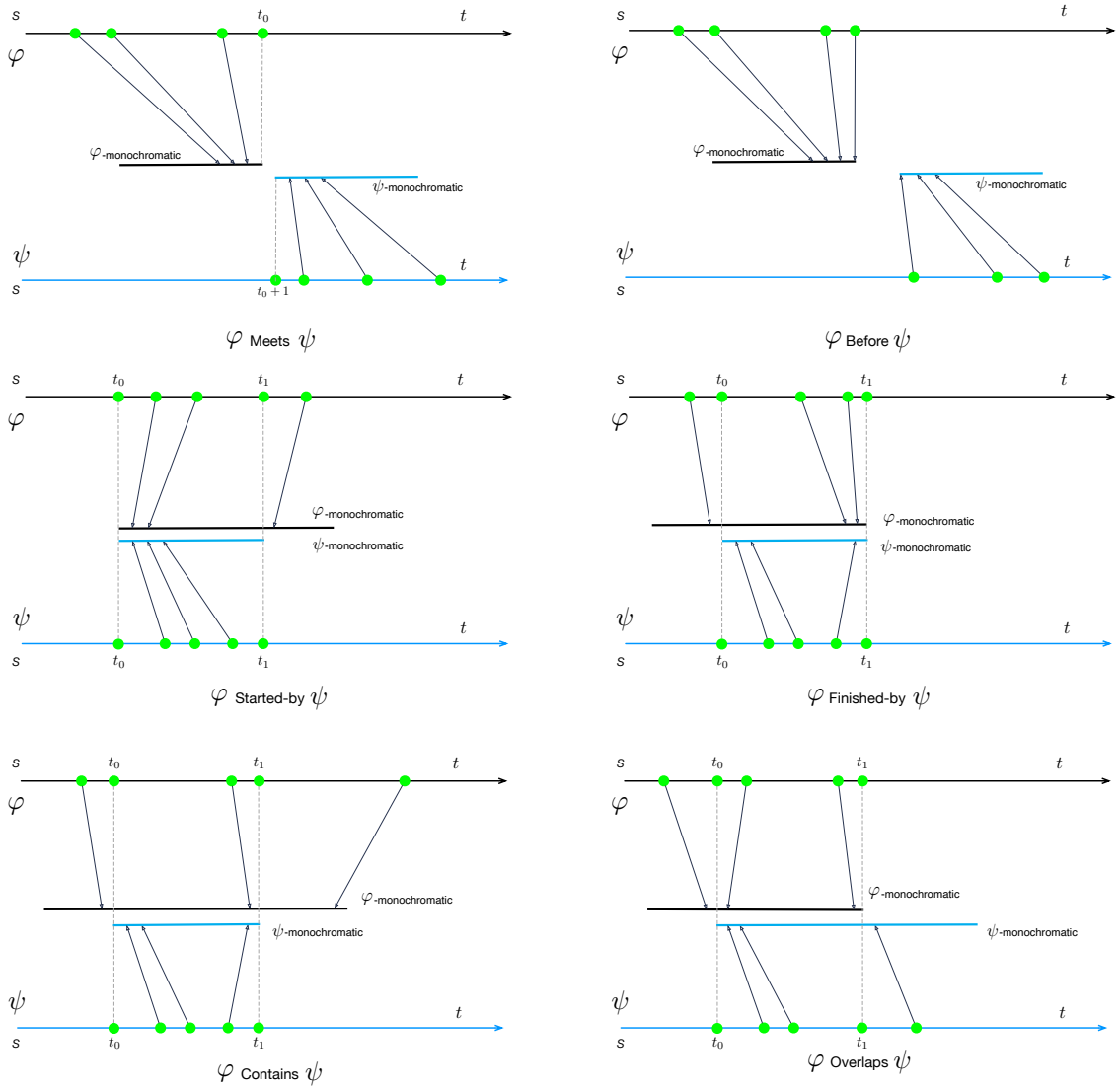

Figure 3: The intended semantics of the TCL modal operators can be understood as Allen's interval relations on the monochromatic subgraphs induced by the respective TCL formulas. Both t_0 and t_1 are displayed in the diagrams, although in some cases only one time point is explicitly mentioned in the definition, and the existence of the other time point is implied.

3.3 Satisfiability, Validity, Equivalence, and Notions of Cohort

With the preparation in the previous sections, we can define the common logical terms in this context. A TCL formula φ is said to be *satisfiable*, if there exists an assignment σ such that $(\sigma, t) \models \varphi$ for some t. A TCL formula φ is said to be *valid*, if for all assignment σ and all t, we have $(\sigma, t) \models \varphi$. We write $\models \varphi$ when φ is valid.

Two TCL formulas φ and ψ is said to be equivalent if for all σ and all t, $(\sigma, t) \models \varphi$ if and only if $(\sigma, t) \models \psi$. We can see that φ and ψ are equivalent if and only if both $\varphi \to \psi$ and $\psi \to \varphi$ are valid.

With these terminology and notations, we can see that all formulas that are valid in the classical proposition logic are also valid for TCL. For temporal operators, our TCL set up entails that, for example, $\models (\varphi \,\mathsf{B}\, \psi) \to (\varphi \,\mathsf{D}\, \psi)$ for any φ and ψ. That is, "started-by" is a special case of "contains."

Our set up allows different specifications of a cohort. The *(global) cohort* defined by φ is the collection of assignments σ such that $(\sigma, 1) \models \varphi$, i.e. $[\![\varphi]\!] := \{\sigma \mid (\sigma, 1) \models \varphi\}$. A *$t$-sectional cohort* is defined as the collection of assignments σ such that $(\sigma, t) \models \varphi$, i.e. $[\![\varphi]\!]_t := \{\sigma \mid (\sigma, t) \models \varphi\}$. To be practically useful, additional time constraints and logical enrichments may be needed (see Results and Discussions).

In biomedical applications such as those exemplified in the next section, finding $[\![\varphi]\!]$ or $[\![\varphi]\!]_t$ against RWD represents a key computational challenge. Compared to classical model-checking for an individual model, we are concerned with

model-checking at a massive scale ("massive model-checking"): the sizes of the resulting cohort ($[\![\varphi]\!]$) can range from thousands to millions, and the background "search space" to be covered can be orders of magnitude larger.

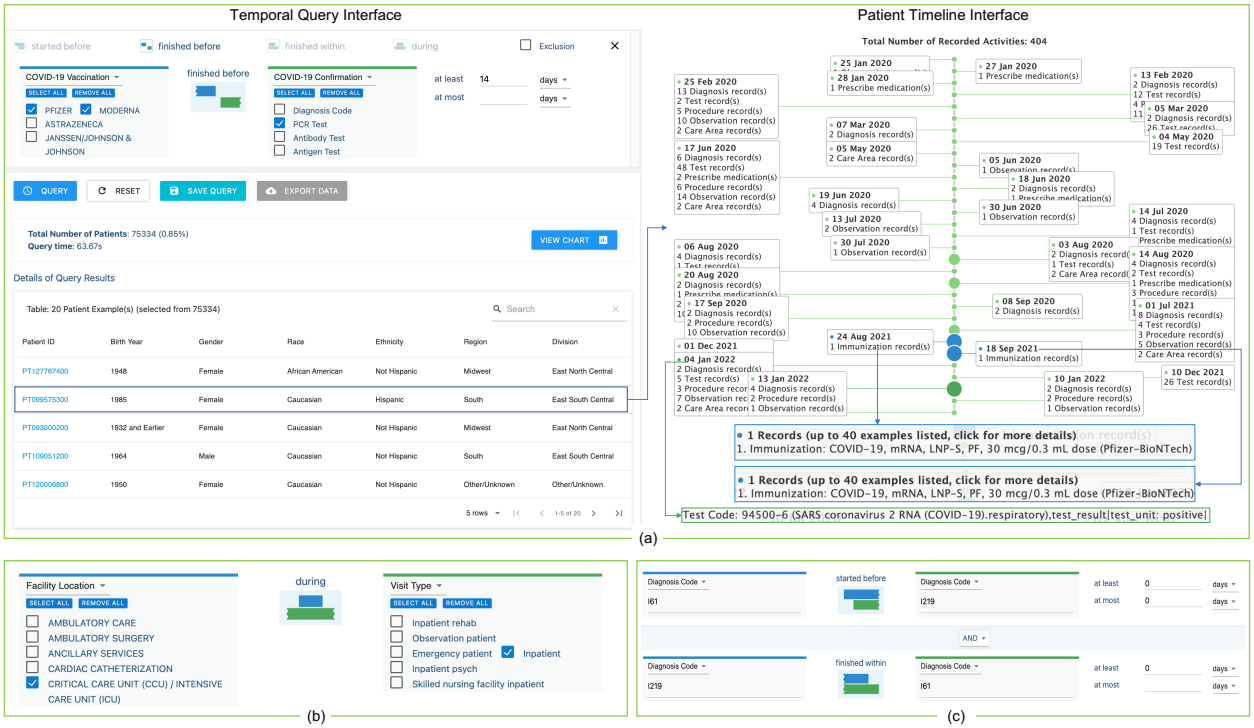

Figure 4: Three temporal query examples with different temporal modalities in CovidSphere.

4 Experimental Results

4.1 Temporal Cohort Discovery in EHR

We demonstrate two use cases where the temporal query interfaces are specified and translated using TCL temporal modalities. In the first use case, we developed a query system called CovidSphere to support a temporal query interface for exploring cohorts within a large COVID-19 EHR dataset. The system uses OPTUM®'s longitudinal EHR data, drawn from dozens of healthcare providers in the United States that include more than 700 hospitals and 7,000 clinics. It included EHR records for 7 million unique individuals who had documented clinical care with a documented COVID-19 encounter or acute respiratory illness after 02/01/2020 and/or documented COVID-19 testing regardless of their results. The main purpose of CovidSphere is to construct and issue queries, especially temporal queries, using a web-based interface without requiring knowledge about how the backend data are structured and stored, thereby shortening the data access life-cycle and facilitating data exploration by researchers.

In CovidSphere, we designed and implemented query widgets to provide four different primitives to formulate different TCL formulas, including started before (i.e., event A started before event B started), finished before (i.e., event A finished before event B started), finished within (i.e., event A finished before event B finished), and during (i.e., event A happened during event B). A user can select different settings for each primitive: 1) clinical events with query criteria; 2) temporal relationship between two query terms, and 3) the time interval between two events that occur in a particular time sequence. All TCLs can be represented by a logical combination of these four primitives.

Three temporal query examples with different TCL modalities are shown in Figure 4. (a) Find patients who were first diagnosed with COVID-19 by PCR testing 14 days of their most recent COVID-19 vaccination. On the top left is the temporal query interface, including the query widgets and results part, and on the right is the visualization of the patient timeline for validating the temporal query results for a given patient. The query results included 75,334 patients, and we manually verified the correctness of the results by visualizing the patient timeline. For example, for patient "PT099575300," after the second shot of COVID-19 vaccine on September 18, 2021, the test result for COVID-19 was positive on January 4, 2022. (b) Find patients who had been in the ICU during their hospitalization. (c) Find patients who had intracerebral hemorrhage (I63) with heart attack (I219) at the same time, with intracerebral

hemorrhage continued after heart attack. We also manually verified query results for (b) and (c) for correctness.

To speed up temporal queries for CovidSphere, a collection of inverted indices were introduced [27]. To date, CovidSphere's temporal query interface has facilitated data extraction for a dozen projects, the results of several of which appeared in peer-reviewed publications [28, 29, 30].

4.2 Neurological Application

Our second use case is SeizureSphere, a web-based graphical query interface for electroencephalographic (EEG) data generated from the Center for SUDEP (Sudden Unexpected Death in Epilepsy) Research (CSR; NIH U01NS090407, U01NS090408). The CSR is a National Institute for Neurological Disorders and Stroke (NINDS) funded "Center Without Walls" consortium initiative for prospective collaborative research in the epilepsies. Researchers from 14 institutions across the U.S. and Europe and brought together extensive and diverse expertise for the unraveling of SUDEP phenomenology, pathophysiology, and biomarker identification. The CSR's Informatics and Data Analytics Core (IDAC; NIH U01NS090408) has so far prospectively collected high grade multimodal data sets including high-resolution EEG signal data sets, sleep PSGs, research-grade brain MRI, biochemical and DNA samples together with detailed phenotypic data [31].

SeizureSphere's backend consisted of a MySQL database to store EEG signals from 1,976 patients with 3,334,298 clinical-grade annotations. SeizureSphere's frontend was a web-based temporal query interface for selecting subgroups of patients with special EEG patterns and characteristics of interest. This interface supports the specification of temporal operators for seizure-related events. The most common temporal modalities for EEG annotation queries are "contains" and "before." "Meets," "started-by" and "finished-by" are not heavily used because of the clinical significance of time granularity involved. Note that although the temporal resolution of clinical-grade EEG data can reach sub-millisecond level, the annotation of clinically meaningful events remains at seconds level.

A screenshot of the graphical temporal query interface for CSR EEG annotations is displayed in Figure 5. The query interface allows a user to add intervals with start and end annotations and drag them into desirable locations to match the intended temporal relations. The temporal modality of this example involves a GTC which consists of a Tonic Phase and a Clonic Phase, and the Tonic Phase must occur *before* the Clonic Phase. Additional time constraint can be added for more complex queries, such as requiring EEG suppression to occur within 1 minute after the end of GTC. SeizureSphere has been used for supporting EEG data extraction for "Cardiac and Autonomic Pathological Markers for Arrhythmias and Sudden Unexpected Death in Epilepsy Patients," a project funded by Citizen's United for Research in Epilepsy (CURE) foundation.

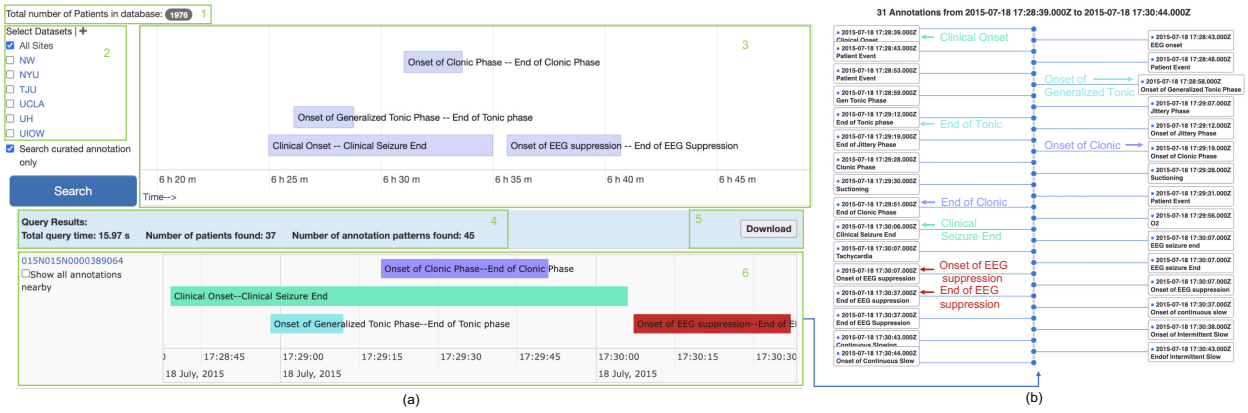

Figure 5: SeizureSphere: a graphical temporal query interface for EEG events. a) A screenshot of the query page, which consists of six components: 1) total number of patients; 2) Epilepsy Monitoring Unit (EMU) site selection; 3) temporal query widgets; 4) results summary; 5) export results to a CSV file; 6) query result with all the matched query terms highlighted in different colors. b) Complete annotation records near the matched pattern for patient "015N015N0000389064." We verified the query results manually by reviewing the matched patterns with the patient's annotation timeline. For example, on July 18, 2015, this patient was observed a tonic phase followed by a clonic phase during a seizure and an EEG suppression within 1 minute after the end of seizure, which matches both the query widget and annotation timeline.

5 Discussion

Unlike typical computer science applications, EHR data, and RWD in general, are "observational" in nature. This entails that such data need to be interpreted using an Open World Assumption. OWA does not make a commitment about the truth status of a statement when its truth cannot be determined based on information in an EHR. As such, RWD can only "approximate" information about reality (e.g., healthcare activities of a person as a representation of actual health status). For example, the date of initial diagnosis of a disease can not be interpreted as the precise starting point of the disease. However, date for lab test has a more precise temporal meaning, due to its "interventional" nature. For such reasons, cohort discovery and development systems serve only as a starting point to refine study subgroups, and TCL and the like serve to support this process.

Although more complex in nature, boolean operation and even nesting of temporal modalities can be essential in some biomedical applications. For example, in an EMU discharge summary report, the section on epileptic seizure semiology many capture a temporal sequence of distinct seizure events such as *Autonomic seizure* $\Rightarrow$ *Right verse seizure* $\Rightarrow$ *Generalized tonic-clonic seizure*, where *Generalized tonic-clonic seizure*, is defined as "generalized tonic phase" followed by a "clonic phase."

We highlighted only two application domains in the Results section, but other applications are clearly possible. For example, in precision medicine, genes, exons, and regulatory regions are annotated subsequences with a chromosome ID with a start and an end position. In this case, the positions do not have a time interpretation, although the sequential nature of whole genomes makes them suitable for a "discrete, linear time" treatment. Indeed, Luo and his group have used Allen's interval algebra for performing efficient genomic interval queries [32], which can serve as an attractive basis for a formal TCL formulation.

Further theoretical developments invite themselves. These include deductive proof systems for TCL, decidability and complexity of TCL fragments, and enriched formulation (multi-value, fuzzification, or timed variants). Naturally, these are topics beyond the scope of this initial paper.

6 Conclusion

This paper introduced Temporal Cohort Logic to fill a conceptual gap in formalizing temporal reasoning in biomedicine. Applications in EHR-based cohort discovery and in neurophysiological data resource have been demonstrated. TCL provides a formal logical framework for reasoning about time in biomedicine, allowing general investigation into the properties of this framework independent of a specific query language or a database system. It also puts our approach in the context of logical developments in computer science and opens opportunities for further translational research into theory and applications.

Acknowledgement. This work was supported in part by the National Science Foundation through grant IIS-2047001 and the National Institutes of Health grants R01LM013335, R01NS126690, R01NS116287, U01NS090408, and U01NS090407. The views of the paper are those of the authors and do not reflect those of the funding agencies.

References

[1] Randhawa G. S, Soltysiak M. P, El Roz H, Souza C. P de, Hill K. A, and Kari L. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: Covid-19 case study. *Plos one*. 2020; 15(4): e0232391.

[2] Alimadadi A, Aryal S, Manandhar I, Munroe P. B, Joe B, and Cheng X. Artificial intelligence and machine learning to fight covid-19. 2020 Apr 1;52(4):200–2.

[3] Henry J, Pylypchuk Y, Searcy T, and Patel V. Adoption of electronic health record systems among us non-federal acute care hospitals: 2008–2015 [onc data brief no. 35]. Office of the National Coordinator for Health Information Technology website.

[4] Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). Journal of the American Medical Informatics Association. 2010 Mar 1;17(2):124–30.

[5] Tao S, Cui L, Wu X, and Zhang GQ. Facilitating cohort discovery by enhancing ontology exploration, query management and query sharing for large clinical data repositories. In *AMIA Annual Symposium Proceedings*; volume 2017. American Medical Informatics Association, 2017; p. 1685.

[6] Ganslandt T, Mate S, Helbing K, Sax U, and Prokosch H. Unlocking data for clinical research–the german i2b2 experience. *Applied clinical informatics*. 2011; 2(01): 116–117.

[7] J. F. Allen. Maintaining knowledge about temporal intervals. Communications of the ACM, 26(11):832-843, Nov. 1983.

[8] Prior A. Past, present and future. *Revue Philosophique de la France Et de l.* 1967; 157.

[9] Blackburn P, De Rijke M, and Venema Y. *Modal logic: graph. Darst*; volume 53. Cambridge University Press; 2002.

[10] Pnueli A. The temporal logic of programs. *18th Annual Symp. Foundations of Computer Science*, IEEE, 1977; p. 46–57.

[11] Stirling C. *Modal and temporal logics*. LFCS, Department of Computer Science, University of Edinburgh; 1991.

[12] Clarke E. M. Model checking. In *International Conference on Foundations of Software Technology and Theoretical Computer Science*. Springer, 1997; p. 54–56.

[13] Vardi M. Y. Model checking for database theoreticians. *International Conference on Database Theory*, 2005; p. 1–16.

[14] Moszkowski B, Manna Z. Reasoning in interval temporal logic. InWorkshop on Logic of Programs 1983 Jun 6 (pp. 371–382). Springer, Berlin, Heidelberg.

[15] Rozier KY. Linear temporal logic symbolic model checking. Computer Science Review. 2011 May 1;5(2):163–203.

[16] Reynolds M. An axiomatization of full computation tree logic. The Journal of Symbolic Logic. 2001 Sep;66(3):1011–57.

[17] Hennessy M, Milner R. Algebraic laws for nondeterminism and concurrency. Journal of the ACM, 1;32(1):137–61, 1985.

[18] Alur R, Henzinger TA. A really temporal logic. Journal of the ACM, 2;41(1):181–203, 1994.

[19] J. Halpern and Y. Shoham. A propositional modal logic of time intervals. Journal of The ACM, 38:935–962, 1991.

[20] Atlas. https://www.ohdsi.org/atlas-a-unified-interface-for-the-ohdsi-tools.

[21] Dobbins N. J, Spital C. H, Black R. A, Morrison J. M, De Veer B, Zampino E, Harrington R. D, Britt B. D, Stephens K. A, Wilcox A. B, et al. Leaf: an open-source, model-agnostic, data-driven web application for cohort discovery and translational biomedical research. *Journal of the American Medical Informatics Association*. 2020; 27(1): 109–118.

[22] Samra H, Li A, and Soh B. Gene2d: A nosql integrated data repository of genetic disorders data. In *Healthcare*; volume 8. Multidisciplinary Digital Publishing Institute, 2020; p. 257.

[23] Zhang GQ, Siegler T, Saxman P, Sandberg N, Mueller R, Johnson N, Hunscher D, and Arabandi S. VISAGE: a query interface for clinical research. *Summit on translational bioinformatics*. 2010; p. 76.

[24] Zhang GQ, Cui L, Lhatoo S, Schuele S. U, and Sahoo S. MEDCIS: multi-modality epilepsy data capture and integration system. In *AMIA Annual Symposium Proceedings*; volume 2014. American Medical Informatics Association, 2014; p. 1248.

[25] Cui L, Zeng N, Kim M, Mueller R, Redline S, and Zhang GQ. X-search: an open access interface for cross-cohort exploration of the national sleep research resource. *BMC medical informatics and decision making*. 2018; 18(1): 1–10.

[26] Callahan A, Polony V, Posada J. D, Banda J. M, Gombar S, and Shah N. H. ACE: the advanced cohort engine for searching longitudinal patient records. *Journal of the American Medical Informatics Association*. 2021; 28(7): 1468–1479.

[27] Huang Y, Li X, Zhang GQ. ELII: A novel inverted index for fast temporal query, with application to a large Covid-19 EHR dataset. Journal of Biomedical Informatics. 2021 May 1;117:103744.

[28] Pérez C. A, Zhang GQ, Li X, Huang Y, Lincoln J. A, Samudralwar R. D, Gupta R. K, and Lindsey J. W. Covid-19 severity and outcome in multiple sclerosis: Results of a national, registry-based, matched cohort study. *Multiple Sclerosis and Related Disorders*. 2021; 55: 103217.

[29] Kim Y, Khose S, Abdelkhaleq R, Salazar-Marioni S, Zhang GQ, and Sheth S. A. Predicting in-hospital mortality using d-dimer in covid-19 patients with acute ischemic stroke. *Frontiers in Neurology*. 2021:12.

[30] Kim Y, Li X, Huang Y, Kim M, Shaibani A, Sheikh K, Zhang GQ, and Nguyen T. P. Covid-19 outcomes in myasthenia gravis patients: Analysis from electronic health records in the united states. *Frontiers in Neurology*. 2022:548.

[31] Li X, Tao S, Lhatoo S, Cui L., Huang Y. and Zhang GQ. A multimodal clinical data resource for personalized risk assessment of sudden unexpected death in epilepsy. *Frontiers in Big Data*. 2022, In Press.

[32] Mao C, Eran A, Luo Y. Efficient Genomic Interval Queries Using Augmented Range Trees. Scientific reports. 2019 Mar 25;9(1):1-2.