Temporal models of care sequences for the exploration of medico-administrative data

Johanne Bakalara

Temporal models of care sequences for the exploration of medico-administrative data

Johanne Bakalara[1,2]

[1] Univ Rennes, Inria, CNRS, IRISA, France
[2] Univ Rennes, EA-7449 REPERES, France

Abstract. Pharmacoepidemiology with medico-administrative databases enables to study health products impact in real-life settings. However, querying this data in a medical perspective is still a challenge because of the complexity of temporal constraints for medical patterns. This PhD thesis aims at proposing a formal framework for the analysis of temporal medical patterns that would support a well-founded tool for querying care trajectories. This paper presents our problematic and introduce a use-case which illustrates the comparison of several querying formalisms.

Keywords: Temporal logics · Ontologies · pharmacoepidemiology.

1 Introduction

Pharmacoepidemiology studies the benefits and risks of drugs usage on population in real life settings. As such, pharmacoepidemiology deals with positive impact as well as safety concerns. Patients with health events are identified according to their individual characteristics, their treatments and their concomitant treatments. Collecting information to answer one specific epidemiological question requires a lot of time and is very expensive. Using medico-administrative databases is an attractive potential alternative because of their large population coverage and their availability. For instance, the SNDS[3] database contains individual information of French patients: age, sex, location; and health reimbursement information: drug deliveries, medical acts or medical visits and hospitalisations (date of arrival, leaving date, diagnosis code) but it does not contain medical reports. The interest of using this medico-administrative database has been demonstrated by the suspension of benfluorex [19].

Epidemiologists use the SNDS database to find patients that experienced some medical events of interest. All the medical events that are stored in the database compose the *care trajectory* of a patient. For epidemiologists, the challenge is to define selection criteria that would reconcile those actual patient information with the medical semantic. The definition of these criteria composes a health pattern called *care pathway*. For example, the care trajectory of a patient "delivery at date t of anticoagulant with ATC code *B01AF01*" would match the

[3] SNDS: French National System for Health Data (previously SNIIRAM).

care pathway "patients exposed to anticoagulant for 3 to 6 months under the assumption of a monthly delivery".

The problem that we address is to enable epidemiologists to query the database of care trajectories with care pathways. The complexity is twofold:

- use of ontological concepts ("Doppler imagery act"/"anticoagulant"/"vascular specialist"): the code of the medical act performed on a patient is given, but is more precise than the criterion "anticoagulant drug" and symbolic domain knowledge is required to reconcile both. Here ,"Anticoagulant" refers to a class of drugs that is described in the ATC taxonomy.[4]
- use of temporal constraints ("in the week after"/"within 4 months"): the temporal order of cares, numerical duration/delays specifies the temporal organisation of the events.

2 Problem statement

The objective of this work is to propose a formal framework that would support a well-founded and efficient tool for querying care trajectories in the context of pharmacoepidemiology.

Generally speaking, let $\mathcal{T} = (T_i)_{i\in[n]}$ be a set of n care trajectories and φ a care pathway abstract description. φ holds in a care trajectory $T \in \mathcal{T}$, denoted $T \models \varphi$, iff the care trajectory verifies the care pathway. The formalisation problem is threefold, we define:

- a formalism to model care trajectories, T, which represent the SDNS data
- a formalism to model care pathways, φ, which specifies an abstract care pathway
- a computational model that can evaluate whether T entails φ: $T \models \varphi$.

As we noticed in the introduction, specifying care pathways requires to manipulate: temporal concepts (time constraints and time window), medical concepts and knowledge (ontologies). The ideal formal framework should capture these dimensions, enable intuitive queries to be expressed for a wide range of pharmacoepidemiological studies and be computationally efficient.

It is of paramount importance to base choices on solid theoretical foundations. Expressiveness and efficiency are known to be antagonist objectives [12]. A well-founded approach would be the basis for proposing long-term solutions, make possible future improvements and facilitate its application to a broad range of contexts (*i.e.*, various databases, queries).

3 Related works

This section presents four families of formalisms related to the problem: model checking; Complex Event Processing (CEP); temporal databases; and Knowledge

[4] ATC: Anatomical Therapeutic Chemical Classification System.

Representation and Reasoning (KR). The formalism should represent data (care trajectories), query (care pathway) and compute the answers of the queries on data. The last two families have been further explored in medical context [8] than the two others.

Model checking [7] verifies if a model satisfies a property or a formula. This research line is interested in representing dynamic systems with formal temporal formalisms (discrete event models, $\mathcal{M}$, describing how the system evolves) to prove some properties specified by formula φ. A formula φ is true iff φ is true for any trace that can be generated from model $\mathcal{M}$. The most common formalisms for timed formula in Model Checking are LTL, CTL or MTL (temporal LTL).

To apply such methods in our context, the events of care trajectory are represented by one finite trace of the system (and there is no system model in our case) and the care pathway is represented by a timed logic formula. The care trajectory is selected iff the trace satisfies the formula. These methods are interesting because they provide formal results (expressiveness, completeness, equivalences) on the representation of timed systems, but they neither manage reasoning nor ontological representation. In the medical domain, model checking has been used to study the compliance of care pathways [5].

Complex Event Processing (CEP) [11] is a research line that aims at processing log-streams with patterns. Log-streams are streams or sequences of timed events. The CEP processes these logs to detect or to locate complex events (or *patterns*) defined by the user. This domain defines formalisms that aim at being very efficient to process streams and expressive to specify patterns. Temporal constraint networks [6] or Chronicles [10] are simple temporal models that are interesting for their graphical representation, but are limited to simple events. Some more complex formalisms, *e.g.* ETALIS [1] or logic-based event recognition [11] propose very expressive representations of complex events, including reasoning techniques (encompassing ontologies) which enrich the CEP capabilities.

In our context, care trajectories are logs, and care pathways are the complex events. We are not interested in the stream dimension, but their formalisms to represent complex events can be adapted in the context of static logs.

Temporal databases [18] extend the notion of database to timestamped data. Databases cover data representation problems but also specific querying language problems. This family encompasses the temporal extension of relational databases (*e.g.* TSQL) but also Semantic Web approaches which combine query languages (*e.g.* SPARQL) and expressive description languages. Care trajectories are facts in the temporal database and the querying of care pathways becomes a problem of specifying care pathways in the query language. [17] shown that the Semantic Web is a relevant approach for our problem, but does not explicitly address the problem of timed queries.

Finally, **Knowledge Representation and Reasoning** (KR) [12] is "the study of how what we know can at the same time be represented as comprehensibly as possible and reasoned with as effectively as possibly". In this research domain, temporal KR is focused on representing and reasoning about time. It gives rise to several logics [13], for instance: Allen's logic, McDermott's

logic, Event Calculus or Halpern & Shoham's logic. KR is a general framework to study how to represent care trajectories and how to model reasoning-based queries on care pathways. Approaches from the other families may be represented with appropriate logics. Studying KR formalisms seems of paramount importance as it provides common foundations to compare various approaches. Description Logic (DL) [3] is a KR formalism allowing ontology-mediated query answering (OMQA) [4]. Artale et *al.* [2] present a temporal extension of DL that may be suitable for our problem. For instance, [16] developed a tool based on OWL for research data management with a temporal reasoning in a clinical trial system.

4 Comparison of approach on a use case

4.1 Rational

This section introduces a real use case. In this example, pharmacoepidemiologists want to select patients with Venous Thromboembolism (VTE) from the data contained in the SNDS. Venous thromboembolism, *i.e.* deep vein thrombosis (DVT) or pulmonary embolism (PE), is a frequent and potentially fatal disease [9, 14]. The motivation is to survey how many people are concerned, if the number of patients increased and if a specific drug has an impact. The difficulty for epidemiologists lies in the description of the care pathways that will accurately identify VTE from the SNDS data. The description below describes two care pathways that physicians proposed to identify VTE.

> *In clinical practice facing a clinical suspicion of VTE, physicians first prescribe anticoagulant and then confirm or not the diagnosis through specific medical acts: for instance Doppler ultrasonography or CT scan. Patients with suspected PE are often hospitalized whereas patients with suspected DVT are managed on a ambulatory basis. If the suspicion is confirmed, anticoagulant deliveries continues for 3 to 12 months or sometimes longer duration. Hence, diagnosis (through medical act) is preceded or followed by anticoagulant initiation within a time window of at most 1 to 2 days, keeping in mind that PE suspicion leads to hospitalisation during which medical act to confirm the diagnosis are performed and then anticoagulant is observed only after the patient comes back home.*

Through these observations, pharmacoepidemiologists identified the following two care pathways to detect patients with VTE from SNDS data:

1. A diagnosis (DVT or PE) or a medical act (Doppler or CT scan) during or prior to anticoagulant (AC) deliveries for 1 to 2 days and delivery lasts a minimum of 3 months and a maximum of 12 months (sometimes longer). Each delivery is separated by 0 to 2 months.
2. A diagnosis PE during an hospitalisation followed by AC delivery.

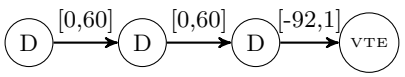
These care pathways are ordered and also contain time constraints between events (for instance number of days) or duration of events (time window). Searching for such patterns requires high expressiveness that make databases query

$B01(\mathsf{Pierre}, n_1)$ $B01(\mathsf{Pierre}, n_2)$
$PE(\mathsf{Pierre}, n_3)$ $B01AF01(\mathsf{Paul}, n_1)$
$B01AF01(\mathsf{Paul}, n_3)$ $EDQM001(\mathsf{Paul}, n_4)$
$n_1 < n_2 < n_3 < n_4$

Fig. 1. ABox

$B01AF01 \sqsubseteq B01AF$
$B01AF \sqsubseteq B01$
$B01 \sqsubseteq B$

Fig. 2. TBox



$(\Diamond D \wedge \bigcirc(\Diamond D) \wedge \bigcirc(\Diamond D)) \wedge (\Diamond(DVT \vee PE))$

$\Diamond(DVT \vee PE) \Rightarrow ((\Diamond_{[0\ 24]} D) \wedge (\Diamond_{[0\ 60]} D) \wedge (\Diamond_{[0\ 60]} D) \wedge (\Diamond_{\geq 365}\ (\Box(\neg D))))$

Fig. 3. Example of care pathway representation of the case study in three formalisms. Top left: Chronicles, top right: LTL formula and bottom: MTL formula (D denotes an anticoagulant delivery event).

languages (SQL, TSQL) practically difficult to use. The next section illustrates formalisms that will allow the generalization of the representation of care pathways and care trajectories.

Note that this use case illustrates the problem of formalizing care pathways of patients suffering from VTE. But, for sake of generality, our formalism has to specify the care pathway patterns of a broad range of care studies.

4.2 Comparison of different models

This section illustrates several formalisms: Description Logic (DL) to represent care trajectories and Chronicles, LTL and MTL to express queries. These queries represent care pathways. We discuss their expressivity in our use case.

DL is used to describe and to reason about concepts on data. The first one consists in defining the data and the data form (called *ABox*) which contains knowledge at the instance level: a set of assertion defining concepts, roles and a countably infinite set of individuals names. Concepts with individuals names and roles with individuals names are forming atoms.

The second step is to define a base of knowledge (*TBox*) which is a set of concepts inclusions. For instance, the concept $B01AF01$ (Rivaroxaban) representing anticoagulant drugs is the leaf concept in the hierarchy of concept modeling the ATC classification: $B01AF$ (Direct Xa inhibitor), $B01$ (Antithrombotic agents) and B (Blood and blood forming organs).

Considering the CCAM (medical acts) code for the Doppler: $EDQM001$ (iliac and lower limb arteries), we could construct the *ABox* of Fig. 4.2, where Pierre and Paul are patients and n_i are dates of medical events. *TBox* of Fig. 4.2 is issued from the previous ATC classification.

The third step is to define a query to extract information from knowledge contained in *ABox* and *TBox*. Usually, queries are expressed with a first order logic. [2] designed a temporal DL: TQL that extends the standard ontology language: OWL 2QL. However, to express queries with temporal delays, several

formalisms seem to be adapted. For example, the chronicle formalism represents a care pathway as a temporal constraint graph. Chronicles allow the expression of sequential order of events with temporal constraints such as interval of time. Furthermore, negative time in the interval expresses that an event may occur before or after another one. Fig. 3 specifies patients having at least three AC deliveries separated by 0 to 60 days, and a diagnosis DVT before, after or during deliveries. DVT occurs 92 days earlier or one day after the third delivery. However, we can not explicitly restrict the number of deliveries to 12 months as defined in the use case. We also can not use the notion of *no event* (event does not occur).

Model checking offer the possibility to express *no event*. Such as an example we propose the LTL formula Fig. 3 as an example applied to our case of study: The LTL formula represents a care pathway with at least three deliveries and a diagnosis DVT or PE. We literally read it: *in the future ($\Diamond$), there is the delivery of B*01 *and ($\wedge$) it is followed ($\bigcirc$), in the future, by the delivery of B*01 *and it is followed, in the future, by a delivery of B*01 *and, in the future, there is a DVT (or $\vee$) PE) diagnosis*. LTL only contains ordered events and does not contain time constraints. The expressivity of such formula is too limited for our problemThe MTL formula adds the capability to express quantitative temporal constraints. Fig. 3 is a MTL formula which represents a care pathway with a DVT or PE followed by three AC deliveries separated between 0 to 60 days, and no deliveries occur after 365 days. MTL can explicitly restrict the number of deliveries and temporal constraints but the notion of sequences is manually found by the multiple use of $\Diamond$.

From a computational aspect, chronicle recognition is a space/time-efficient task. Simple LTL formula would also be space/time-efficient to check but it is expressively poor. In contrast, MTL is more expressive but it is known to be undecidable. This is a theoretical limitation but not necessarily a practical constraint [15].

5 Conclusion

Current researches are based on comparing temporal models and on studying their suitability for pharmacoepidemiology studies. Participating to this doctoral consortium would be an opportunity to discuss existing temporal model in medical applications and to identify other similar use cases that will highlight other needs. Expectation from the doctoral consortium is to guide me for the future evolution of my thesis.

Acknowledgments This multidisciplinary PhD thesis began six months ago. It is supervised by the pharmacoepidemiologist E. Oger in the REPERES team and by computer scientists, T. Guyet and O. Dameron in the IRISA laboratory, and A. Happe in the REPERES team.

References

1. Anicic, D., Fodor, P., Rudolph, S., Stühmer, R., Stojanovic, N., Studer, R.: ETALIS: Rule-based reasoning in event processing. In: Proc. Reasoning in event-based distributed systems, pp. 99–124 (2011)
2. Artale, A., Kontchakov, R., Kovtunova, A., Ryzhikov, V., Wolter, F., Zakharyaschev, M.: Ontology-mediated query answering over temporal data: A survey. In: nternational Symposium on Temporal Representation and Reasoning (TIME) (2017)
3. Baader, F., Calvanese, D., McGuinness, D., Patel-Schneider, P., Nardi, D.: The description logic handbook: Theory, implementation and applications. Cambridge university press (2003)
4. Bienvenu, M.: Ontology-mediated query answering: harnessing knowledge to get more from data. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI). pp. 4058–4061 (2016)
5. Bottrighi, A., Giordano, L., Molino, G., Montani, S., Terenziani, P., Torchio, M.: Adopting model checking techniques for clinical guidelines verification. Artificial intelligence in medicine **48**(1), 1–19 (2010)
6. Cabalar, P., Otero, R.P., Pose, S.G.: Temporal constraint networks in action. In: ECAI. pp. 543–547 (2000)
7. Clarke Jr, E.M., Grumberg, O., Kroening, D., Peled, D., Veith, H.: Model checking (2018)
8. Combi, C., Keravnou-Papailiou, E., Shahar, Y.: Temporal information systems in medicine (2010)
9. Delluc, A., Ianotto, J.C., Tromeur, C., De Moreuil, C., Couturaud, F., Lacut, K., Le Moigne, E., Louis, P., Thereaux, J., Metges, J.P., et al.: Real-world incidence of cancer following a first unprovoked venous thrombosis: Results from the EPIGETBO study. Thrombosis research **164**, 79–84 (2018)
10. Dousson, C., Le Maigat, P.: Chronicle recognition improvement using temporal focusing and hierarchization. In: Proc. IJCAI. pp. 324–329 (2007)
11. Giatrakos, N., Artikis, A., Deligiannakis, A., Garofalakis, M.: Complex event recognition in the big data era. In: Proc. VLDB Endow. vol. 10, pp. 1996–1999 (2017)
12. Levesque, H.J.: Knowledge representation and reasoning. Annual review of computer science **1**(1), 255–287 (1986)
13. Long, D.: A review of temporal logics. The Knowledge Engineering Review **4**(2), 141–162 (1989)
14. Oger, E., EPI-GETBO: Incidence of venous thromboembolism: a community-based study in western france. Thrombosis and haemostasis **83**(05), 657–660 (2000)
15. Ouaknine, J., Worrell, J.: On the decidability of metric temporal logic. In: 20th Annual IEEE Symposium on Logic in Computer Science (LICS'05). pp. 188–197 (2005)
16. O'Connor, M.J., Shankar, R.D., Parrish, D.B., Das, A.K.: Knowledge-data integration for temporal reasoning in a clinical trial system. International journal of medical informatics **78**, 77–85 (2009)
17. Rivault, Y., Dameron, O., Le Meur, N.: queryMed: Semantic Web functions for linking pharmacological and medical knowledge to data. Bioinformatics (2019)
18. Snodgrass, R.T.: Temporal databases. In: Proc. IEEE computer (1986)
19. Weill, A., Païta, M., Tuppin, P., Fagot, J.P., Neumann, A., Simon, D., Ricordeau, P., Montastruc, J.L., Allemand, H.: Benfluorex and valvular heart disease: a cohort study of a million people with diabetes mellitus. Pharmacoepidemiology and drug safety **19**(12), 1256–1262 (2010)