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

Johanne Bakalara, Thomas Guyet, Olivier Dameron, Emmanuel Oger, André Happe

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

Johanne Bakalara[1,2], Thomas Guyet[1], Olivier Dameron[1]
Emmanuel Oger[2,3], André Happe[2,4]

[1] Univ Rennes, Inria, CNRS, IRISA
[2] Univ Rennes, EA-7449 REPERES
[3] CHU Rennes
[4] CHRU Brest
johanne.bakalara@irisa.fr

Abstract : Pharmacoepidemiology with medico-administrative databases enables to study impact of health products in real-life setting. These studies require to manipulate the raw data, the care trajectories, in order to identify pieces of data that may witness the medical information that is looked for. The manipulation can be seen as a querying process in which a query is a description of a medical pattern (*e.g.* occurrence of illness) with the available raw features from care trajectories (*e.g.* occurrence of medical procedures, drug deliveries, etc.). The more expressive is the querying process, the more accurate is the medical pattern search. The temporal dimension of care trajectories is a potential information that may improve the description of medical patterns. The objective of this work is to propose a formal framework that would design a well-founded tool for querying care trajectories with *temporal medical patterns*. In this preliminary work, we present the problematic and we introduce a use case which illustrates the comparison of several querying formalisms.

Mots-clés : Temporal logics, Description logic, Ontologies, Pharmacoepidemiology, SNDS.

1 Introduction

Pharmacoepidemiology studies uses and effects of drugs on population in real life including benefits and risks. As such, pharmacoepidemiology deals with positive impact (*i.e.* prevention of disease) 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 epidemiological question requires a lot of time and is very expensive.

Medico-administrative databases are a potential alternative which are attractive because of their large population coverage and their availability. For instance, the SNDS[1] (Tuppin *et al.*, 2017) 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 (Weill *et al.*, 2010). It was the first large-scale pharmacoepidemiology study in France that was possible thanks to information contained in the SNDS.

The study shows that diabetic patients exposed to Benfluorex have an higher risk of hospitalisation for heart disease than the unexposed diabetics in the following years. Epidemiologists use the SNDS database to find patients that experienced some medical events of interest. It is worth noticing that information has been collected for administrative purposes (care reimbursements), but not for medical ones. As a consequence, the medical content associated with medical events is relatively poor.

[1]SNDS: French National System for Health Data (previously SNIIRAM). The SNDS is the world largest medico-administrative database with a population coverage close to 99%.

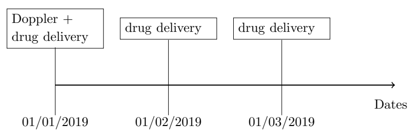

Figure 1: Example of a care trajectory.

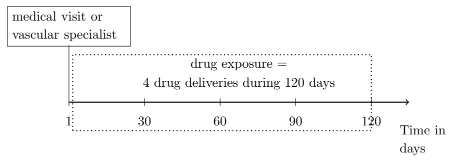

Figure 2: Example of a care pathway.

Let us consider a patient having a Venous Thromboembolism (VTE). The database does not contain the information that *the patient has the disease VTE*. This information has to be deduced from the information about its care deliveries. For instance, a consultation with a vascular specialist and deliveries of anticoagulant drugs. These pieces of information are available in the data in a structured format. Compared to medical record in hospital, the SNDS database has less medical information, but it does not required sensitive text analysis.

All medical events that are stored in the database compose the care trajectory of a patient. Figure 1 illustrates a care trajectory. The challenge of the epidemiologist 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 illustrated Figure 2. A possible care pathway of patients with VTE[2] is a consultation with a **vascular specialist** or at the hospital with an **Doppler imagery act** (CCAM[3] *EDQM001* code) followed by at least 3 deliveries of **anticoagulant drugs** **within 4 months** (with the first delivery of anticoagulant **in the week after** consultation/Doppler). The care trajectory denotes the actual medical events stored in the database while a care pathway is a pattern of medical events that describe the pathology profile as illustrated on Figures respectively 1 and 2.

We aim at enabling epidemiologists to query the database of care trajectories with such kind of complex descriptions of care pathway. The complexity is twofold:

- use of ontological concepts (Doppler imagery act/anticoagulant/vascular specialist): the code of the medical act is given, but the code for anticoagulant drugs is not precisely given. Anticoagulants refer 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.

Our work focuses on having expressive temporal constraints in the description of care pathways. Temporal constraints enable to discriminate care trajectories of interest from care trajectories with same events but presenting different delays (*e.g.* an Doppler imagery act followed by an anticoagulant delivered several weeks later does not witness a VTE). This may help to specify care pathways that match only the desired care trajectories. It is worth noting that drugs deliveries, medical acts and hospital stays are available in the database with timestamps.

The challenge we face is summed up in Figure 3. Care pathways are on the left side and care trajectories are on the right side. The overall objective is to bridge the two sides. On the top in red, epidemiologists formulates a medical hypothesis. In blue, at the second level, epidemiologist describes care pathways with the available health information and a database stores health information. In green, at the third level, data and queries are specified in a formalism. This latter has to express a maximum of information to represent the complexity of the data (care trajectories) and of the care pathway. Green and blue levels are intertwined. Today, the tools for querying SNDS (third level) lack of expressiveness and constraint epidemiologists for their specification (SQL, SAS, etc...).

[2]This description is an illustration for pedagogical purposes, we refer to the use case for a medical description of VTE.

[3]CCAM: Classification Commune des Actes Médicaux

[4]ATC: Anatomical Therapeutic Chemical Classification System.

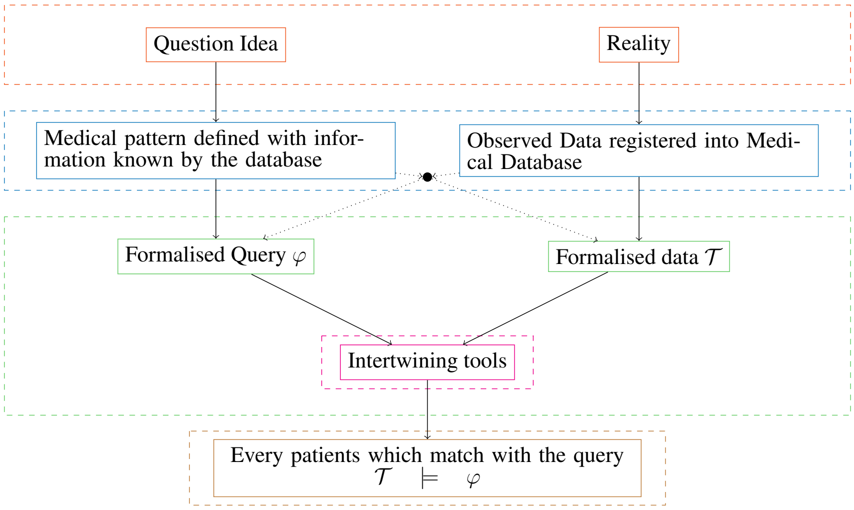

Figure 3: From medical study issue to the care trajectory querying

In this article, we show that the problem of querying patient care trajectories with temporal medical patterns (the care pathways) may be addressed, sometimes only partially, by tools or formalisms coming from different computer science fields: complex event processing, formal reasoning, knowledge representation and database. In this preliminary work, our objective is to identify strengths and weaknesses of these approaches.

In the next sections, we states formally of problem of querying a set of care pathways, then Section 3 reviews related work. The last section illustrates some of approaches of the related work on a case study.

2 Problem statement

The objective of this work is to propose a formal framework that would design 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 $T \models \varphi$, if and only if the care trajectory *contains* the care pathway. The formalisation problem is threefold:

- define a formalism to model care trajectories, T, which represent the SDNS data
- define a formalism to model care pathways, φ, which specifies an abstract care pathways
- define 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 (Levesque, 1986). 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 work

This section presents four families of formalisms to answer the problem: model checking; Complex Event Processing (CEP); temporal databases; and Knowledge Representation and Reasoning (KR). The work that we mentioned addresses the three problems mentioned in Section 2 at the same time. The formalism should represent data (care trajectories), query (care pathway) and to compute the answers of the queries on data. The last two families have been more explored in medical context (Combi *et al.*, 2010) than the two others.

Model checking (Clarke Jr *et al.*, 2018) 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 if and only if φ is true for any traces [5] that can be generated from the model $\mathcal{M}$. The most common formalisms for formula in Model Checking are LTL (Linear Temporal Logic), CTL (computation tree logic) or MTL (metric temporal lgic) which is the temporal extension of 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 pathways is represented by a temporal logic formula. The care trajectory is selected if and only if 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 don't manage neither reasoning nor ontological representation. In related medical domain, model checking has been used to study the compliance of care pathways (Bottrighi *et al.*, 2010).

Complex Event Processing (CEP) (Giatrakos *et al.*, 2017) 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 (Cabalar *et al.*, 2000) or Chronicles (Dousson & Le Maigat, 2007) are simple temporal models that are interesting for their graphical representation, but are limited to simple relational events. Some more complex formalisms, *e.g.* ETALIS (Anicic *et al.*, 2011) or logic-based event recognition (Giatrakos *et al.*, 2017), propose very expressive representations of complex events, including reasoning techniques (including ontologies) which enrich the capabilities of CEP.

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

Temporal databases (Snodgrass, 1992) extend the notion of database to timestamped data. Databases issued data representation problems but also specific querying language problems. We gather in this family the temporal extension of relational databases (*e.g.* TSQL) but also web semantic 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. Rivault *et al.* (2019) shown that semantic web is an interesting approach for our problem, but does not explicitly address the problem of timed queries.

Finally, **Knowledge Representation and Reasoning** (KR) (Levesque, 1986) 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 (Long, 1989), for instance: Allen's logic, McDermott's logic, Event Calculus or Halpern & Shoham's logic.

[5] Sequences taht register the set of atomic propositions that are valid along the execution, *cf.* part 3.2.2 in (Baier & Katoen, 2008)

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) (Baader *et al.*, 2003) is a KR formalism allowing ontology-mediated query answering (OMQA) (Bienvenu, 2016). Artale et *al.* (Artale *et al.*, 2017) present a temporal extension of DL that may be suitable for our problem. For instance, (O'Connor *et al.*, 2009) 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 (Oger & EPI-GETBO, 2000; Delluc *et al.*, 2018). This requires 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 (referring to Figure 3, this description is between the red and blue levels).

> *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 0 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 form SNDS data (referring to Figure 3 this description is in the blue level):

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 contain sequential order but also 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 languages (SQL, TSQL) practically difficult to use. Next section illustrates alternative solutions to represent the green level (see Figure 3).

This use case illustrated 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 as many case studies as possible.

4.2 Comparison of different models

This section aims at illustrating the comparison of four formalisms, Description Logic (DL), Chronicles, LTL and MTL to discuss about their power of expressivity in our case of study.

DL supplies ontologies query answering technologies for Ontology-Based Data Access (OBDA) and is well known to represent medical data. Indeed, Description Logic is used to describe and reason about concepts on data. Reasoning with Description Logic is performed in three steps. 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 consists in define a base of knowledge (*TBox*) which is a set of concepts inclusions. Concept inclusions represent a hierarchy of concepts. For instance, the concept $B01AF01$ representing anticoagulant drugs is the leaf concept in the hierarchy of concept modeling the ATC classification:

- B: Blood and blood forming organs
  - B01: Antithrombotic agents
    - B01AF: Direct Xa inhibitor
      - B01AF01: Rivaroxaban

And considering the CCAM (medical acts) code for the Doppler: EDQM001 (iliac and lower limb arteries) we could construct the following *ABox*, where Pierre and Paul are patients and n_i are dates of medical events:

$$B01(\mathsf{Patient1}, t_1) \quad B01(\mathsf{Patient1}, t_2) \quad PE(\mathsf{Patient1}, t_3)$$
$$B01AF01(\mathsf{Patient2}, t_1) \quad B01AF01(\mathsf{Patient2}, t_3) \quad EDQM001(\mathsf{Patient2}, t_4)$$
$$t_1 < t_2 < t_3 < t_4$$

And *TBox* issued from the previous ATC classification:

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

$B01AF01$ is the subclass of $B01AF$ which is the subclass of $B01$ which is the subclass of B. Operators linking concepts are defined depending on the class of DL chosen. For example, the $\mathcal{ALC}$ family offers the concept constructors: negation, conjunction, disjunction, existential restriction $\exists$ and universal restriction $\forall$.

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. Artale *et al.* (2017) designed a temporal DL: TQL that extends the standard ontology language: OWL 2QL. It offers the capability of having time as individuals names and to compare whether an atom occurs before another. In our case study, we propose a query to find patients designed by the use case defined in Section 4.1.

$$\exists patientID, t_1(t_{ref} < t_1) \wedge EDQM001(patientID, t_{ref}) \wedge B01(patientID, t_1)$$
$$\wedge t_2(t_2 < t_1) \wedge B01(patientID, t_2)$$
$$\wedge t_3(t_3 < t_1) \wedge B01(patientID, t_3)$$

We find patients having three deliveries of $B01$ (and subclasses of $B01$) in a row and the first one is delivered after an Doppler. The temporal information that can be represented here is limited to the order of the events. Delay or duration can not be specified.

To express queries with temporal delays, the formalism of chronicle 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

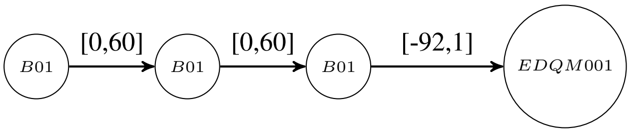

Figure 4: Chronicles

interval expresses that an event may occur before or after another one. Figure 4 specifies patients having at least three anticoagulant 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 cannot use the notion of *no event* (event does not occur). Model checking offer the possibility to express *no event* and can be used as queries. Such as an example we propose the following LTL formula as an example applied to our case of study:

$$(\Diamond D_{B01} \wedge \bigcirc(\Diamond D_{B01}) \wedge \bigcirc(\Diamond D_{B01}) \wedge (\Diamond(DVT \vee PE))$$

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 $B01$ and ($\wedge$) it is followed ($\bigcirc$), in the future, by the delivery of $B01$ and it is followed, in the future, by a delivery of $B01$ and, in the future, there are the diagnosis DVT or ($\vee$) PE).* LTL only contains order between events and doesn't contains time constraints. It is quite limited for our problem, so we refer to its temporal extension MTL. The MTL formula adds the capability to express quantitative temporal constraints. We propose the following MTL formula as an example applied to our case of study:

$$\Diamond(DVT \vee PE) \Rightarrow ((\Diamond_{[0\ 2]} D_{BO1}) \wedge (\Diamond_{[0\ 60]} D_{BO1}) \wedge (\Diamond_{[0\ 60]} D_{BO1}) \wedge (\Diamond_{\geq 365}\,(\Box(\neg D_{BO1})))$$

It represents a care pathway with a DVT or PE followed between 0 to 2 days after 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 $\Diamond$. For deep understanding of notations, we refer the reader to Bouyer *et al.* (2005).

From computational point of view, Chronicles may be very space/time-efficiently to be recognizes in care trajectories. Simple LTL formula would also be space/time-efficient to check but it is expressively poor. In contrast, MTL is known to be undecidable. It is a theoretical limitation but, not necessary a practical constraint Ouaknine & Worrell (2005).

5 Conclusion

In this article, we introduced the context of pharmaco-epidemiological studies with medico-administrative databases and the challenge to query such databases with medical questions. It consists in finding patients satisfying a medical pattern that we want to be expressive.

To express medical patterns, we use the formalism of Description Logic to describe data and include medical knowledge issued from the available classification (*e.g.* ATC, CCAM). To query these data, we compared solutions issued from first order logic, Chronicles and MTL. It shows that none of them is enough expressive to correctly *translate* all the desirable constraints of the care pathway. The future work consists of extending these formalisms and to study efficiency issues by testing them through many cases of study on the SNDS.

References

ANICIC D., FODOR P., RUDOLPH S., STÜHMER R., STOJANOVIC N. & STUDER R. (2011). ETALIS: Rule-based reasoning in event processing. In *Proceedings of the Conference on Reasoning in event-based distributed systems*, p. 99–124.

ARTALE A., KONTCHAKOV R., KOVTUNOVA A., RYZHIKOV V., WOLTER F. & ZAKHARYASCHEV M. (2017). Ontology-mediated query answering over temporal data: A survey. In *Proceedings of the International Symposium on Temporal Representation and Reasoning (TIME)*, p. 1–37.
BAADER F., CALVANESE D., MCGUINNESS D., PATEL-SCHNEIDER P. & NARDI D. (2003). *The description logic handbook: Theory, implementation and applications*. Cambridge university press.
BAIER C. & KATOEN J.-P. (2008). *Principles of model checking*. MIT press.
BIENVENU M. (2016). Ontology-mediated query answering: harnessing knowledge to get more from data. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, p. 4058–4061.
BOTTRIGHI A., GIORDANO L., MOLINO G., MONTANI S., TERENZIANI P. & TORCHIO M. (2010). Adopting model checking techniques for clinical guidelines verification. *Artificial intelligence in medicine*, **48**(1), 1–19.
BOUYER P., CHEVALIER F. & MARKEY N. (2005). On the expressiveness of tptl and mtl. In *International Conference on Foundations of Software Technology and Theoretical Computer Science*, p. 432–443: Springer.
CABALAR P., OTERO R. P. & POSE S. G. (2000). Temporal constraint networks in action. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, p. 543–547.
CLARKE JR E. M., GRUMBERG O., KROENING D., PELED D. & VEITH H. (2018). *Model checking*. Springer.
COMBI C., KERAVNOU-PAPAILIOU E. & SHAHAR Y. (2010). *Temporal information systems in medicine*. Springer.
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.* (2018). Real-world incidence of cancer following a first unprovoked venous thrombosis: Results from the EPIGETBO study. *Thrombosis research*, **164**, 79–84.
DOUSSON C. & LE MAIGAT P. (2007). Chronicle recognition improvement using temporal focusing and hierarchization. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, p. 324–329.
GIATRAKOS N., ARTIKIS A., DELIGIANNAKIS A. & GAROFALAKIS M. (2017). Complex event recognition in the big data era. *Proceedings of Conference on Very Large Data Base Endowment (VLDB Endow.)*, **10**(12), 1996–1999.
LEVESQUE H. J. (1986). Knowledge representation and reasoning. *Annual review of computer science*, **1**(1), 255–287.
LONG D. (1989). A review of temporal logics. *The Knowledge Engineering Review*, **4**(2), 141–162.
OGER E. & EPI-GETBO (2000). Incidence of venous thromboembolism: a community-based study in western france. *Thrombosis and haemostasis*, **83**(05), 657–660.
OUAKNINE J. & WORRELL J. (2005). On the decidability of metric temporal logic. In *20th Annual IEEE Symposium on Logic in Computer Science (LICS'05)*, p. 188–197.
O'CONNOR M. J., SHANKAR R. D., PARRISH D. B. & DAS A. K. (2009). Knowledge-data integration for temporal reasoning in a clinical trial system. *International journal of medical informatics*, **78**, 77–85.
RIVAULT Y., DAMERON O. & LE MEUR N. (2019). queryMed: Semantic web functions for linking pharmacological and medical knowledge to data. *Bioinformatics*, p. to appear.
SNODGRASS R. T. (1992). Temporal databases. In *Theories and methods of spatio-temporal reasoning in geographic space*, p. 22–64. Springer.
TUPPIN P., RUDANT J., CONSTANTINOU P., GASTALDI-MÉNAGER C., RACHAS A., DE ROQUEFEUIL L., MAURA G., CAILLOL H., TAJAHMADY A., COSTE J. *et al.* (2017). Value of a national administrative database to guide public decisions: From the système national d'information interrégimes de l'assurance maladie (sniiram) to the système national des données de santé (snds) in france. *Revue d'epidemiologie et de sante publique*, **65**, S149–S167.
WEILL A., PAÏTA M., TUPPIN P., FAGOT J.-P., NEUMANN A., SIMON D., RICORDEAU P., MONTASTRUC J.-L. & ALLEMAND H. (2010). Benfluorex and valvular heart disease: a cohort study of a million people with diabetes mellitus. *Pharmacoepidemiology and drug safety*, **19**(12), 1256–1262.