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

Supervisors:

- Thomas Guyet, IRISA-Inria/LACODAM (thomas.guyet@irisa.fr)
- Pr. Emmanuel Oger, PU/PH, Head of EA-7449 REPERES
- Olivier Dameron, IRISA-Inria/DYLISS
- Andr Happe, EA-7449 REPERES

Hosting laboratories (in Rennes/France):

- IRISA-Inria/LACODAM Team: `https://team.inria.fr/lacodam/`
- CHU-EHESP REPERES Team: `https://www.ea-reperes.com/`

Funding: Funding the Ph.D. fellowship is funded for 3 years (MESR).
Doctoral schools: MathSTIC and Bio/Health

Keywords: temporal models, sequences, temporal queries, care pathways

Pharmacoepidemiology is the study of the use of drugs under real conditions. Ongoing opening of access of the medico-administrative databases is a scientific breakthrough in this medical research field. Medico-administrative databases contain data collected for administrative purposes. The French SNDS[1] (previously SNIIRAM) is the world largest medico-administrative database with a coverage close to 99% of the population. This makes this database a real treasure for both epidemiologists and data scientists. These data (drug deliveries, medical consultations, hospitalization, in fact all health care services reimbursed by Health Insurance) constitute a wealth of readily available information. Its use in pharmacoepidemiology responds to the need for rapid answers to public health questions. However, addressing both the quantity and complexity of health data remains an open challenge.

The difficulty of analyzing medico-administrative data is the semantic gap between the raw data (for example, database record about the delivery at date t of drug with ATC[2] code $N02BE01$) and the nature of the events sought by clinicians ("was the patient exposed to a daily dose of paracetamol higher than $3g$?"). The solution that is used by epidemiologists consists in enriching the data with new types of events that, on the one side, could be generated from raw data and on the other side, have a medical interpretation. Such new abstract events are defined by clinician using *proxies*. For example, drugs deliveries can be translated in periods of drug exposure (drug exposure is a time-dependent variable for non-random reasons) or identify patient stages of illness, etc. A *proxy* can be seen as an abstract description of a care sequence.

Currently, the clinicians are limited in the expression of these *proxies* both

[1] SNDS: Systme National des Donnes de Sants

[2] ATC: Anatomical Therapeutic Chemical Classification System

by the coarse expressivity of their tool and by the need to process efficiently large amount of data. [6] From a semantic point of view, care sequences must fully integrate the **temporal** and taxonomic dimensions of the data to provide significant expression power. From a computational point of view, the methods employed must make it possible to efficiently handle large amounts of data (several millions care pathways).

The aim of this thesis is to study temporal models of sequences in order 1) to show their abilities to specify complex *proxies* representing care sequences needed in pharmaco-epidemiological studies and 2) to build an efficient querying tool able to exploit large amount of care pathways.

In previous works, we focused on the chronicle model [5, 7] which represents a care sequence as a set of events for which numerical constraints are added on the delay between their occurrences. One advantage of this simple model is that it could be easily visualized by clinicians. In addition, it is effective for querying large masses of sequences but shows limits in its expressiveness (especially on taxonomies or the expression of disjunctions). Other models of behavior have been proposed with different time models coming from various communities (*e.g.* logic [2, 8], discrete event systems [3, 9] or automatic verification [1]). Each of these representations therefore offers higher semantic power but also computational limits (decidability, efficiency, etc).

This thesis will contribute to the PEPS plateform [4] developed in collaboration by IRISA and REPERES. Querying tools based on temporal models will be deployed and evaluated on real pharmacoepidemiological studies in close relationship with epidemiologists. Model expressivity will be evaluated according to the practical needs of clinicians both from theoretical and practical points of view.

The main stages of the PhD thesis will be: 1) state of the art, discovery of the SNDS and pharmaco-epidemiology, 2) identify potential models of care sequences and selection of 2 to 4 typical pharmacoepidemiology studies to reproduce, 3) implement, evaluate and compare temporal models and 4) valorize the work through studies and publications.

Candidate profile:

- preferably student preparing or having MSc diploma (master 2) within one of this specialities:
  - MSc Diploma in theoretical computer science (algorithmics, logic or formal models, data science, artificial intelligence) with strong interest in medical application and abilities to work in this application field
  - MSc Diploma in (bio)medical informatics with good backgrounds in computer science
- good abilities to work in a multidisciplinary environment
- good communication skills in English (oral and written)
- autonomy and motivation for research

- good programming skills (knowledge in Python or C++ will be appreciate)
- basic knowledge in logic

Application (by **May 15, 2018**):
Applications must be send to thomas.guyet@irisa.fr including:

- detailed CV,
- motivation letter explaining your interest for this subject,
- MSc transcripts with your rank among your peers (last available transcript and course syllabus for final year MSc),
- MSc internship information (and thesis, if available),
- contacts for recommendation [optional].

Some interviews will be offered between May 28 and June 4. The final decision will be given in June. The PhD thesis is expected to start in September (or October) 2018.

References

[1] Sundararaman Akshay, Loïc Hélouët, Claude Jard, and Pierre-Alain Reynier. Robustness of time petri nets under guard enlargement. In *International Workshop on Reachability Problems*, pages 92–106. Springer, 2012.

[2] James F Allen and George Ferguson. Actions and events in interval temporal logic. *Journal of logic and computation*, 4(5):531–579, 1994.

[3] Alexander Artikis, Anastasios Skarlatidis, François Portet, and Georgios Paliouras. Logic-based event recognition. *The Knowledge Engineering Review*, 27(4):469–506, 2012.

[4] F. Balusson, M.-A. Botrel, O. Dameron, Y. Dauxais, E. Drezen, A. Dupuy, T. Guyet, D. Gross-Amblard, A. Happe, N. Le Meur, B. Le Nautout, E. Leray, E. Nowak, C. Rault, E. Oger, and E. Polard. PEPS: a platform for supporting studies in pharmaco-epidemiology using medico-administrative databases. In *Proceedings of international Congress on e-Health Research*, 2016.

[5] Yann Dauxais, Thomas Guyet, David Gross-Amblard, and André Happe. Discriminant chronicles mining. In *Proceedings of Conference on AIME*, pages 234–244, 2017.

[6] Erwan Drezen, Thomas Guyet, and André Happe. From medico-administrative databases analysis to care trajectories analytics: An example with the french SNDS. *Fundamental & Clinical Pharmacology*, 2017.

[7] Houssam-Eddine Gougam, Yannick Pencolé, and Audine Subias. Diagnosability analysis of patterns on bounded labeled prioritized petri nets. *Discrete Event Dynamic Systems*, 27(1):143–180, 2017.

[8] Erik T Mueller. Event calculus. *Foundations of Artificial Intelligence*, 3:671–708, 2008.

[9] Ariane Piel, Jean Bourrely, Stéphanie Lala, Sylvain Bertrand, and Romain Kervarc. Temporal logic framework for performance analysis of architectures of systems. In *NASA Formal Methods Symposium*, pages 3–18. Springer, 2016.