From medico-administrative databases analysis to care trajectories analytics: an example with the French SNDS

Erwan Drezen, Thomas Guyet, André Happe

From medico-administrative databases analysis to care trajectories analytics: An example with the French SNDS

Erwan DREZEN [1,3], Thomas GUYET [2], André HAPPE [1,4]
[1] Plateforme PEPS/ANSM, UPRES EA 7449 REPERES
[2]AGROCAMPUS-OUEST/IRISA – UMR6074
[3] CHU de RENNES, [4] CHRU de BREST

Abstract

Medico-administrative data like SNDS (*Système National de Données de Santé*) are not collected initially for epidemiological purpose. Moreover, the data model and the tools proposed to SNDS users make difficult their in-depth exploitation. We propose a data model, called the ePEPS model, based on health care trajectories to provide a medical view of raw data. A data abstraction process enables the clinician to have an intuitive medical view of raw data and to design a study-specific views. This view is based on a generic model of care trajectory, *i.e.* a sequence of timestamped medical events for a given patient. This model is combined with tools to manipulate efficiently care trajectories.

1. Introduction

Medico-administrative databases hold rich information about health care trajectories (or health care pathways) at an individual level. Such data is very valuable for carrying out pharmaco-epidemiological studies on large representative cohorts of patients in real life conditions. Moreover, historical data is readily available for longitudinal analysis of care trajectories. These opportunities are given by the use of the database of the French health care system, so called SNDS database, which covers 98.8% of the French population with a sliding period of 3 years.

A classical pharmaco-epidemiological study from medico-administrative databases consists in three main steps: 1) defining inclusion and exclusion criteria of a cohort, 2) specifying proxies for events of interest and 3) analyzing the transformed data. Practically, these three steps are closely intertwined and make use of digital data management tools (*e.g.* SQL databases, R or SAS). The study outcomes depend on the available data at hand as much as on the tools to manage and process them.

But the data model[1], designed for administrative purposes, is not suitable for pharmaco-epidemiological studies without careful data preparation. It leads to difficulties for epidemiologists to access the useful information and even to know what is reachable with such databases. For instance, the SNDS database is a relational database with hundred of tables with very complex join relations. The set of prescribed drugs of a patient is accessible with a query containing 10 join relations involving attributes with unintuitive names. Mastering the data management with such complex model requires a lot of time, a good knowledge of its content and some technical skills. It is a practical bottleneck to exploit the potential of the database.

Our approach aims at providing a care trajectory view of the raw data. The care trajectory appears to be a natural medical and longitudinal view of individual patient data. This generic representation of the data will empower the clinician to analyze the data. It is more intuitive than complex relational databases and explicitly captures its temporal dimension.

The challenge for computer sciences is to produce tools to abstract the raw data model into a health care trajectory model suitable for queries at a high level of expressiveness. The expected properties of our tools are the following:

- generic: tools have to be used to answer a broad range of studies carried out by epidemiologists
- flexible: the integration of complementary and new data sources must be done easily and transparently
- expressive: we aim at empowering the epidemiologist to analyze more deeply the databases than existing solutions, and more especially on their temporal and ontological dimensions.

2. Abstract model of care trajectories

We developed a two layered approach to abstract the raw databases into study-specific view of patients health care trajectories. The first layer is a generic model that abstracts raw administrative databases into a medical view of the data. It proposes a generic model of care trajectory. The second layer provides a tool to flexibly transform care trajectories and to provide views that are specific to a study.

1 A data model is an abstract model that describes the organization of the data. In relational database, it is the description of tables, their attributes and their relations.

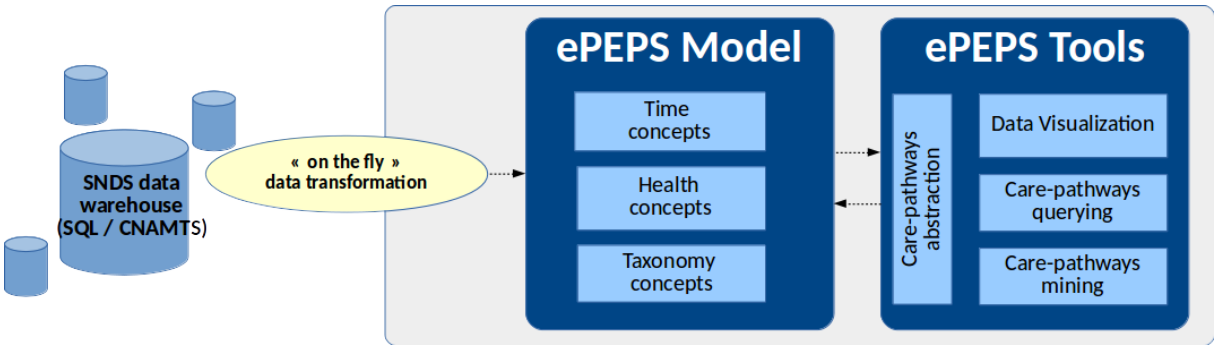

Figure 1: Software components architecture. ePEPS model is constructed on the fly from raw SNDS data. ePEPS tools interact with the data through this model.

The ePEPS model encompasses a generic model of care trajectories and a software component (see Figure 1). The model formalizes the elementary notions of care trajectory with explicit medical concepts (such as the notion of patient, practitioner, drugs deliveries, etc). It provides a longitudinal view of the different events related to an actor (patient or practitioner). In addition, usual taxonomies used in medical field are also fully managed in the ePEPS model. The software component provides a programming interface reifying the model concepts into digital objects that can be used through several widely used programming languages like *java* and *python*. It constitutes the bedrock of our research and development efforts for developing new tools. Through this interface, it is possible to access to any patient events in a single intuitive code line.

The specificity of our approach is to be based on *wrappers* that transform on-the-fly the original SNDS data into care trajectories. This avoids to modify the original data model or to duplicate the data. A wrapper embeds complex SQL queries to access the data. The advantages of this wrappers based approach are three-folds: first, it provides already optimized queries that are hidden to the user. This ensure an efficient access to the data without specific knowledge about the SNDS data model. Second, it enables to link transparently several databases (coming from cohorts or other sources). Once the wrappers have been designed in the new database (describing how to join its content to the main SNDS database) then the ePEPS model gives access to its content in the same manner as for the original data. Third, the approach smooths the SNDS schema changes along time. The SNDS continuously evolves, *e.g.* names of variables or tables. The knowledge of these changes are embedded in the wrappers such that changes are not visible to the end-user.

The second layer of the data abstraction provides a view on care trajectories specific to a study. Final statistical analysis require to define complex transformations of the original data. Abstraction bridges the semantic gap between care trajectories at hand and the required information to carry out a reliable analysis. Transformations include the selection of meaningful events in care trajectories and the definition of new proxies via dedicated algorithms (such as drugs exposure reconstruction from drugs deliveries). Such algorithms may be tedious to set up but are of particular importance for the analysis outcomes. Our care trajectory manipulation tool enables to define such transformations as processing blocks. In addition, we propose a parameterized SQL-like formal language that manipulates care trajectories. In this language, a query specifies a workflow of treatments – the processing blocks – that transforms the data in the ePEPS model into a dataset to be analyzed. Depending on the data analysis tools, it can be tabular datasets or sequences of events, *i.e.* simplified care trajectories. Such design based on treatment blocks and workflow specifications makes easier the modification and the reuse of the usual data transformation algorithms.

3. Chronicles: a tool for analyzing care trajectories

The ePEPS model component provides an intuitive representation of the data as care trajectories. But such complex data can not be directly manipulated by standard data analytic tools (such as SAS or R). This motivates the development of specific tools for care trajectory analysis.

We more especially investigate the chronicle model of a care sequence. Chronicles have been initially proposed in the field of alarm log monitoring[(1)]. It models the behaviors of a system as a set of events constrained by temporal relations. Figure (2) illustrates a chronicle of a simple care sequence: “Three drugs have been delivered: once drug A and twice drug C. The second delivery of C occurs between two and four days before the first ones and the first drug C has been delivered at least 3 days before and at most 1 day after A.”

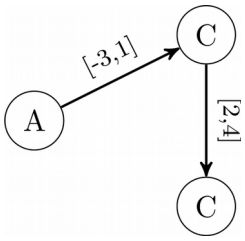

Figure 2: Example of Chronicle

Chronicle is a temporal model that is easy to understand and to specify by clinicians. Moreover, there exists efficient algorithms dedicated to find chronicle matches in long care trajectories and algorithms to extracts frequent and discriminant chronicles[(2)] from a large set of care trajectories. We use this model to enrich the querying and analysis capabilities. Such tools will be integrated in the ePEPS tools.

References

[1] Dousson, C., & Duong, T. V.. Discovering Chronicles with Numerical Time Constraints from Alarm Logs for Monitoring Dynamic Systems. In proceedings of *IJCAI*, 1999, pp. 620-626.

[2] Dauxais Y., Guyet T., Gross-Amblard D., Happe A. Discriminant Chronicles Mining. In: ten Teije A., Popow C., Holmes J., Sacchi L. (eds) Proceedings of Artificial Intelligence in Medicine, 2017.