ARTICLE IN PRESS

Computers in Biology and Medicine ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

# Mining approximate temporal functional dependencies with pure temporal grouping in clinical databases

Carlo Combi [a], Matteo Mantovani [a], Alberto Sabaini [a], Pietro Sala [a], Francesco Amaddeo [b], Ugo Moretti [b], Giuseppe Pozzi [c,*]

[a] Dipartimento di Informatica, Università degli Studi di Verona, strada le Grazie 15, I-37134 Verona, Italy
[b] Dipartimento di Sanità Pubblica e Medicina di Comunità, Università degli Studi di Verona, p.le L.A. Scuro 10, I-37134 Verona, Italy
[c] DEIB, Politecnico di Milano, p.za L. da Vinci 32, I-20133 Milano, Italy

A R T I C L E I N F O

*Article history:*
Received 17 February 2014
Accepted 6 August 2014

*Keywords:*
Approximate temporal functional dependency
Temporal granule
Sliding window
Grouping
Psychiatric patients
Pharmacovigilance

A B S T R A C T

Functional dependencies (FDs) typically represent associations over facts stored by a database, such as "*patients with the same symptom get the same therapy*." In more recent years, some extensions have been introduced to represent both temporal constraints (temporal functional dependencies – TFDs), as "*for any given month, patients with the same symptom must have the same therapy, but their therapy may change from one month to the next one*," and approximate properties (approximate functional dependencies – AFDs), as "*patients with the same symptom* generally *have the same therapy*." An AFD holds most of the facts stored by the database, enabling some data to deviate from the defined property: the percentage of data which violate the given property is user-defined.

According to this scenario, in this paper we introduce approximate temporal functional dependencies (ATFDs) and use them to mine clinical data. Specifically, we considered the need for deriving new knowledge from psychiatric and pharmacovigilance data.

ATFDs may be defined and measured either on temporal granules (e.g. grouping data by day, week, month, year) or on sliding windows (e.g. a fixed-length time interval which moves over the time axis): in this regard, we propose and discuss some specific and efficient data mining techniques for ATFDs. We also developed two running prototypes and showed the feasibility of our proposal by mining two real-world clinical data sets. The clinical interest of the dependencies derived considering the psychiatry and pharmacovigilance domains confirms the soundness and the usefulness of the proposed techniques.



## 1. Introduction

Current clinical database systems enable us to store huger and huger quantities of data, and data mining techniques help in extracting relevant knowledge from these data. Analyzing temporal evolution of data, time series, changes of information over time, may lead to additional *temporal* knowledge. *Temporal* data mining is the research field in this direction, working on structured [1] and, occasionally, on semi-structured data [2].

Knowledge on (clinical) databases may be expressed in two ways: on one hand, it can be represented through suitable constraints on data; on the other hand, it can be derived through the analysis of data by discovering patterns, regularities, and so on.

According to the first point of view and considering data stored in a plain relational database, we may express constraints by identifying functional dependencies (FD). Let us consider, for example, a simple database table describing the reference areas for emergency admissions in a region. We typically specify that patients have a single reference hospital for emergencies, depending on their address (considering the neighborhood of the admitting reference hospital). We can thus specify a functional data dependency between the home address of the patient and the location of the hospital: all patients with the same address must refer to the same hospital. Leveraging the definition of functional dependencies as a way of expressing constraints on data, the research community focused also on extending FDs to deal with data temporalities [3–7]: as example, a temporal functional dependency (TFD) may be used to express the

* Corresponding author.
*E-mail addresses:* Carlo.Combi@univr.it (C. Combi), Matteo.Mantovani@univr.it (M. Mantovani), Alberto.Sabaini@univr.it (A. Sabaini), Pietro.Sala@univr.it (P. Sala), Francesco.Amaddeo@univr.it (F. Amaddeo), Ugo.Moretti@univr.it (U. Moretti), giuseppe.pozzi@polimi.it (G. Pozzi).

http://dx.doi.org/10.1016/j.compbiomed.2014.08.004
0010-4825/

Please cite this article as: C. Combi, et al., Mining approximate temporal functional dependencies with pure temporal grouping in clinical databases, Comput. Biol. Med. (2014), http://dx.doi.org/10.1016/j.compbiomed.2014.08.004

constraint that the reference hospital for emergencies depends on the patient home address, but this dependency may change according to the season of the year.

On the other hand, a different approach has to be taken if we consider, for example, a database table collecting data on patients who were admitted for emergencies to hospitals. In this case, we cannot constrain patient addresses to hospitals in a strict way, but we could discover on the collected data that the dependency between patient addresses and hospitals hold on *most* tuples of the database, but not on *all* the tuples of that database. We call this an *approximate* functional dependency (AFD): patients with the same home address usually go to the same hospital (not always the reference one) when they are at home, but, as an example, some patients on holiday could have been admitted to an hospital which is not the closest one to their home address. The issue of discovering approximate functional dependencies from data has been largely studied in the literature [8–11].

As final consideration, we may also experience that over some periods of the year we generally observe an approximate functional dependency, while in some other periods we observe a different approximate dependency: for example, it could occur that patients go to different hospitals for emergencies even according to some specific skills of hospitals in managing seasonal pathologies. In this case, it still holds that we can discover approximate dependencies between patient addresses and hospitals for emergencies, but only if we group data according to the season and the year of the emergency admission. We call this an *approximate, temporal* functional dependency (ATFD). At the best of our knowledge, studies on *approximate temporal* functional dependencies still lack.

According to the depicted scenario, the aim of this paper is to propose a first step, focusing on a specific type of ATFD, of a general framework for temporal data mining of clinical data. In particular, we adopt a framework for temporal functional dependencies recently proposed by Combi et al. [7]: the framework subsumes all the previous proposals dealing with temporal functional dependencies for relational databases and introduces some new kinds of temporal functional dependencies. According to this framework, we then focus on the issue of mining (approximate) temporal functional dependencies based on a temporal grouping of tuples. We introduce the concept of approximate temporal functional dependency with temporal grouping, and discuss through some examples both the case when grouping is induced by granularities (i.e. time units) and the case when sliding windows are used. Then, we propose efficient algorithms for this kind of temporal data mining. Finally, we discuss the application of our algorithms to real world clinical data from the psychiatric and pharmacovigilance domains.

Besides the technical performances, we discuss the clinical meaning and the most relevant mined temporal dependencies; in this regard, it is worth noting that the mined temporal functional dependencies are a relatively new kind of clinical knowledge on data, which deserves further efforts to become clearly interpretable by physicians in a daily clinical setting. Indeed, while association rules and temporal association rules have been considered in clinical domains for years and their role in the clinical decision-support process has been widely acknowledged [12,13], approximate temporal functional dependencies represent a new piece of knowledge that has to be properly integrated in clinical decision-support processes. As an example, temporal association rules may allow one to derive knowledge as "most patients presenting a symptom of chest pain overlapping nausea receive, within few days, a therapy with acetylsalicylic acid". On the other side, approximate temporal functional dependencies provide knowledge at a higher abstraction level, as "in most cases, patients with the same symptoms are given the same drug (i.e. active principle), considering a time window of 10 days". Such kind of knowledge refers to a general relationship between some features of a patient, in this case symptoms and therapies: the relationship holds for any specific values of such features. Such a valuable kind of knowledge requires physicians to merge it with more specific knowledge, such as that one coming from temporal association rules, in the whole decision making process.

The main novelty aspects of this paper can be summarized as in the following, even with a specific reference to the preliminary work in [14], where the main focus was on the proposal of ATFDs and on some preliminary experiments on a reduced set of psychiatry data with some first prototypal algorithms:

- we discuss in detail the proposed approach for ATFDs and introduce completely new algorithms both for granularity-based temporal mining and for mining through sliding windows;
- we present and discuss two important clinical domains, i.e. psychiatry and pharmacovigilance, where temporal data mining is highly required. As the mined temporal dependencies are sometime completely new and unexpected even to expert physicians, we discuss here some possible interpretations of the discovered knowledge;
- the new experimental results, with a new and extended setting considering two different data sets from psychiatry and pharmacovigilance, consist of both a detailed performance analysis and an evaluation and discussion of the mined ATFDs from a clinical point of view.

In the following, we describe the background and the related work (Section 2) and discuss the two clinical domains we considered for temporal data mining, namely psychiatry and pharmacovigilance (Section 3); we introduce the concept of approximate temporal functional dependency (ATFD), providing some examples on the application scenario (Section 4); then we describe how to mine minimal ATFDs (Section 5) and deploy the proposed techniques in clinical domains; we describe the experimental results obtained by considering data in the two mentioned domains (Section 6), and finally (Section 7) we draw some conclusions and sketch out some possible directions for future research.

## 2. Background and related work

We recall here the definition of functional dependency (FD), and then introduce its extensions: approximate functional dependency (AFD) and temporal functional dependency (TFD). Such concepts will lead to the definition of approximate temporal functional dependency (ATFD) of Section 4, where ATFD inherits the properties both from AFD and from TFD.

The concept of functional dependency (FD) comes from the database theory and is defined as follows [15]:

**Definition 2.1** (*Functional dependency*). Let $r$ be a relationship over the relational schema $R$: let $X, Y \subseteq R$ be attributes of $R$. We assert that $r$ fulfills the functional dependency $X \to Y$ (written as $r \vDash X \to Y$) if the following condition holds: $\forall t, t' \in r(t[X] = t'[X] \Rightarrow t[Y] = t'[Y])$