Discovering Predictive Dependencies on Multi-Temporal Relations

Beatrice Amico, Carlo Combi, Pietro Sala, Romeo Rizzi

Department of Computer Science, University of Verona

Athens, TIME 2023

Table of Contents

Table of Contents

Introduction

Motivation: temporal patterns represent an explainable way to study the intrinsic data dependencies. Mining functional dependencies can be fruitfully exploited to improve prediction, often related to ML models.

Goal: we propose a temporally-oriented data mining framework to support the prediction based on the identification of recurring temporal patterns, the **Approximate Temporal Predictive Functional Dependencies** (APFDs), within a **3-window-based temporal framework**.

Functional dependency

An FD is composed of the antecedent (X) and the consequent (Y). Informally, for all the couples of tuples t and t' showing the same value(s) on X, the corresponding value(s) on Y are identical.

$$X \rightarrow Y$$

Through the use of functional dependencies, we can express concepts such as: *"for each drug with a given symptom the disease does not change"*:

$$Drug, Symptom \rightarrow Disease.$$

Temporal functional dependency

When we add **temporal extensions** to the atemporal functional dependencies, we talk about temporal functional dependency (TFD).

Through the use of temporal functional dependencies, we can express concepts such as *"for each drug with a given symptom the received diagnosis does not change, over a time windows of 10 days"*:

[10 days] Drug, Symptom $\Rightarrow$ Diagnosis

Approximate Functional Dependencies

An AFD f requires the FD to be satisfied by most tuples of relation w. It allows a very small portion of tuples of w to violate the dependency.

If this portion is less than or equal to the satisfaction threshold ε, f is approximately satisfied on s.

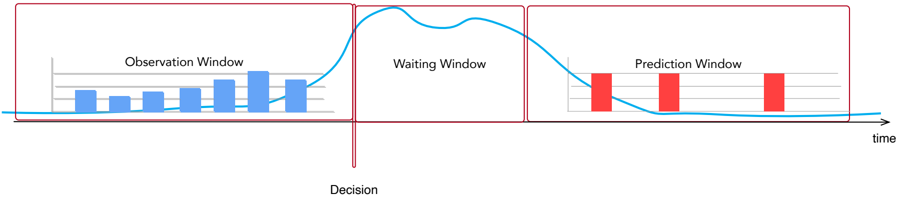
Prediction?

Table of Contents

A 3-window model for the interpretation of predictive temporal data

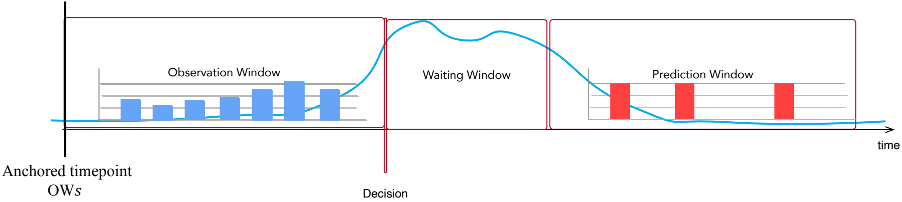
We generalize an approach based on three (possibly moving) time windows:

- **Observation window**: a time interval where the information is collected;
- **Waiting window**: the minimum time interval required to act in order to prevent the event in the prediction window;
- **Prediction window**: the time interval when the predicted event occurs.

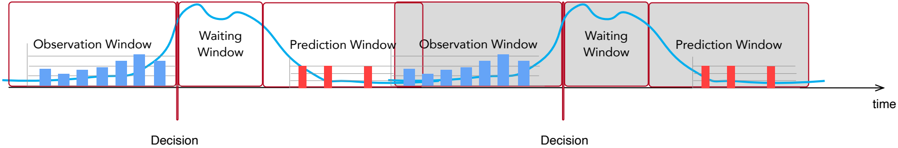


Anchored and unanchored windows

Anchored time windows represent specific periods of the considered time axis.



Unanchored time windows represent windows that "move" through the time axis, constraining only the distance between the considered data.



Fixed and variable length

A second distinction for the time windows, which may provide different results for prediction is:

- **fixed-length**: OW, WW, PW have a fixed length without any further constraint related to the temporal position of data inside them;
- **variable-length**: OW, WW, PW end with the last time point associated with the data to consider in the window.

Multi-temporal relational model

A multi-temporal relation *mrt* is characterized by multiple valid times. Each tuple of such relation represents a piece of history of a given entity, through the values of attributes holding at different (valid) times.

#	**Patient**	$\overline{\text{HR}}^1$	$\overline{\text{VT}}^1$	$\overline{\text{SpO}_2}^2$	$\overline{\text{VT}}^2$	$\overline{\text{Drug}}^3$	$\overline{\text{VT}}^3$	$\dot{\text{AKI}}$	$\dot{\text{VT}}$
1	Daisy	High	19	High	21	Aspirin	23	False	28
2	Daisy	Low	2	High	4	Aspirin	6	False	18
3	Daisy	Low	2	Medium	4	Aspirin	6	False	12
4	Daisy	Medium	5	Medium	7	Indapamide	9	False	18
5	Luke	Low	7	High	8	Ibuprofen	12	True	17
6	Luke	Low	7	High	8	Ibuprofen	12	True	21
7	Luke	Medium	9	High	13	Sulindac	14	True	18
8	Luke	Medium	9	High	13	Sulindac	14	True	21
9	Stevie	Medium	4	Medium	7	Metolazone	8	True	13
10	Stevie	High	1	Low	2	Aspirin	5	False	8
11	Stevie	High	1	Low	2	Indapamide	7	False	8
..	...	...	...	...	...	...	...	...	...
36	Stevie	High	1	Low	2	Aspirin	5	False	25
..	...	...	...	...	...	...	...	...	...

Multi-temporal relational model

A multi-temporal relation *mrt* is characterized by multiple valid times. Each tuple of such relation represents a piece of history of a given entity, through the values of attributes holding at different (valid) times.

#	**Patient**	$\overline{\mathrm{HR}}^1$	$\overline{\mathrm{VT}}^1$	$\overline{\mathrm{SpO_2}}^2$	$\overline{\mathrm{VT}}^2$	$\overline{\mathrm{Drug}}^3$	$\overline{\mathrm{VT}}^3$	$\dot{\mathrm{AKI}}$	$\dot{\mathrm{VT}}$
1	Daisy	High	19	High	21	Aspirin	23	False	28
2	Daisy	Low	2	High	4	Aspirin	6	False	18
3	Daisy	Low	2	Medium	4	Aspirin	6	False	12
4	Daisy	Medium	5	Medium	7	Indapamide	9	False	18
5	Luke	Low	7	High	8	Ibuprofen	12	True	17
6	Luke	Low	7	High	8	Ibuprofen	12	True	21
7	Luke	Medium	9	High	13	Sulindac	14	True	18
8	Luke	Medium	9	High	13	Sulindac	14	True	21
9	Stevie	Medium	4	Medium	7	Metolazone	8	True	13
10	Stevie	High	1	Low	2	Aspirin	5	False	8
11	Stevie	High	1	Low	2	Indapamide	7	False	8
..	...	...	...	...	...	...	...	...	...
36	Stevie	High	1	Low	2	Aspirin	5	False	25
..	...	...	...	...	...	...	...	...	...

Multi-temporal relational model

A multi-temporal relation *mrt* is characterized by multiple valid times. Each tuple of such relation represents a piece of history of a given entity, through the values of attributes holding at different (valid) times.

#	**Patient**	$\overline{\text{HR}}^1$	$\overline{\text{VT}}^1$	$\overline{\text{SpO}_2}^2$	$\overline{\text{VT}}^2$	$\overline{\text{Drug}}^3$	$\overline{\text{VT}}^3$	$\dot{\text{AKI}}$	$\dot{\text{VT}}$
1	Daisy	High	19	High	21	Aspirin	23	False	28
2	Daisy	Low	2	High	4	Aspirin	6	False	18
3	Daisy	Low	2	Medium	4	Aspirin	6	False	12
4	Daisy	Medium	5	Medium	7	Indapamide	9	False	18
5	Luke	Low	7	High	8	Ibuprofen	12	True	17
6	Luke	Low	7	High	8	Ibuprofen	12	True	21
7	Luke	Medium	9	High	13	Sulindac	14	True	18
8	Luke	Medium	9	High	13	Sulindac	14	True	21
9	Stevie	Medium	4	Medium	7	Metolazone	8	True	13
10	Stevie	High	1	Low	2	Aspirin	5	False	8
11	Stevie	High	1	Low	2	Indapamide	7	False	8
..	...	...	...	...	...	...	...	...	...
36	Stevie	High	1	Low	2	Aspirin	5	False	25
..	...	...	...	...	...	...	...	...	...

Multi-temporal relational model

A multi-temporal relation *mrt* is characterized by multiple valid times. Each tuple of such relation represents a piece of history of a given entity, through the values of attributes holding at different (valid) times.

#	**Patient**	$\overline{HR}^1$	$\overline{VT}^1$	$\overline{SpO_2}^2$	$\overline{VT}^2$	$\overline{Drug}^3$	$\overline{VT}^3$	$\dot{AKI}$	$\dot{VT}$
1	Daisy	High	19	High	21	Aspirin	23	False	28
2	Daisy	Low	2	High	4	Aspirin	6	False	18
3	Daisy	Low	2	Medium	4	Aspirin	6	False	12
4	Daisy	Medium	5	Medium	7	Indapamide	9	False	18
5	Luke	Low	7	High	8	Ibuprofen	12	True	17
6	Luke	Low	7	High	8	Ibuprofen	12	True	21
7	Luke	Medium	9	High	13	Sulindac	14	True	18
8	Luke	Medium	9	High	13	Sulindac	14	True	21
9	Stevie	Medium	4	Medium	7	Metolazone	8	True	13
10	Stevie	High	1	Low	2	Aspirin	5	False	8
11	Stevie	High	1	Low	2	Indapamide	7	False	8
..	...	...	...	...	...	...	...	...	...
36	Stevie	High	1	Low	2	Aspirin	5	False	25
..	...	...	...	...	...	...	...	...	...

Time-frame tuple consistency

Given a multi-temporal relation *mtr*, now we are interested in verifying which tuples are "fine" with, or "contained" in, a given time frame.

We are interested in eliciting those tuples having the k observation-related valid times contained in the observation window OW, and the last valid time in the prediction window PW.

We will call them **consistent** with the considered time frame.

Time-frame tuple consistency

Given a tuple t of a multi-temporal relation *mrt*, we say that t is time-frame consistent if the formula $\Theta(t, \alpha, m, [i_1, i_2])$ holds.

There exist different possible formulas according to the different choice of variables:

- α: anchored or unanchored time frame;
- m: fixed or flex modality;
- $[i_1, i_2]$: VT attribute range within the observation window.

Time-frame tuple consistency

Given a tuple t of a multi-temporal relation *mrt*, we say that t is time-frame consistent if the formula $\Theta(t, \alpha, m, [i_1, i_2])$ holds.

There exist different possible formulas according to the different choice of variables:

- α: anchored or unanchored time frame;
- m: fixed or flex modality;
- $[i_1, i_2]$: VT attribute range within the observation window.

$$\Theta(t, \alpha, \text{'}flex\text{'}, [i_1, i_2]) \equiv t[\overline{VT}^{i_2}] - t[\overline{VT}^{i_1}] \leq OW \wedge t[\dot{V}T] - t[\overline{VT}^{i_2}] > WW \wedge t[\dot{V}T] - t[\overline{VT}^{i_2}] < WW + PW$$

Discovering Predictive Dependencies on Multi-Temporal Relations

General idea: propose a general framework allowing the definition of "specialized" functional dependencies having:

- the **antecedent** composed of a set of attributes related to "past" properties, called predictive attributes, ordered according to the corresponding valid times;
- the **consequent** composed of a set of attributes related to "future" properties, called predicted attributes.

Definition

Given:

- an mt-relation schema $MTR(Z\overline{U}^1\overline{U}^2..\overline{U}^k\dot{U} \cup \{\overline{VT}^1, \overline{VT}^2, .., \overline{VT}^k, \dot{VT}\})$ where U_i is a set of attributes representing properties of an entity and Z are the identification attributes;
- a time frame;
- a modality $m \in \{`flex', `fixed'\}$.

a **Predictive Functional Dependency** is expressed as:

$$\boldsymbol{S\overline{P}^h\overline{Q}^i ... \overline{R}^j \xrightarrow[\alpha,m]{} \dot{Y}} \text{ with } 1 \leq h < i < ... < j \leq k$$

where $S \subseteq Z, \overline{P}^h \subseteq \overline{U}^h, \overline{Q}^i \subseteq \overline{U}^i, \overline{R}^j \subseteq \overline{U}^j$ and $\dot{Y} \subseteq \dot{U}$ is the predicted attribute set.

Discovering Approximate PFD (APFD)

We need to deal with some kind of approximation, as it could happen that some PFDs hold on a subset of tuples of the time-frame relation view, we consider.

In other words, we require a PFD f to be satisfied by most tuples of the TF-view w, $w \subseteq mtr$.

A very small portion of tuples of w is allowed to violate the dependency. In the context of APFDs, we consider three error measures: G_3, H_3, J_3.

Approximation: Error G_3

Given a TF-view $w = TFv(mtr, \alpha, m, [1,k])$ of an mt-relation mtr, and a PFD $S\overline{P}^h\overline{Q}^i \ldots \overline{R}^j \xrightarrow[\alpha,m]{} \dot{Y}$, where $S \subseteq Z, \overline{P}^h \subseteq \overline{U}^h, \overline{Q}^i \subseteq \overline{U}^i, \overline{R}^j \subseteq \overline{U}^j$ and $\dot{Y} \subseteq \dot{U}$, and any relation $s \subseteq w$, such that $s \models^E_{\alpha,m} S\overline{P}^h\overline{Q}^i \ldots \overline{R}^j \rightarrow \dot{Y}$, **we define three errors**:

- G_3 considers the minimum number of tuples in w to be deleted to obtain a relation s where the given FD holds.

  G_3 is expressed as:

  $$G_3 = |w| - |s|$$

  The related *scaled measurement* g_3 is defined as:

  $$g_3 = \frac{G_3}{|w|}$$

Approximation: Error H_3

- H_3 is focused on the number of entities that we accept to discard for the sake of the PFD (for example disregard data of entities with a very low number of tuples, which could create noise in our dataset).

  H_3 is expressed as:

$$H_3 = |\{t[Z] \mid \exists t \in w\}| - |\{t[Z] \mid \exists t \in s\}|$$

  The related *scaled measurement* h_3 is defined as:

$$h_3 = \frac{H_3}{|\{t[Z] \mid \exists t \in w\}|}$$

- J_3 considers the number of tuples for each entity we accept to discard to satisfy the *PFD*. It ensures to maintain enough "consistent" information for each entity.

  J_3 is expressed as:

  $$J_3 = \max_{(v \in \{t[Z] | t \in s\})} \{|w_{[v]}| - |s_{[v]}|\}$$

  $w_{[v]} \equiv \{t[Z] | t \in w \wedge t[Z] = v\}$ and $s_{[v]} \equiv \{t[Z] | t \in s \wedge t[Z] = v\}$

  The related *scaled measurement* j_3 is defined as follows:

  $$j_3 = \max_{(v \in \{t[Z] | t \in s\})} \{\frac{|w_{[v]}| - |s_{[v]}|}{|w_{[v]}|}\}$$

Approximate Predictive Functional Dependency (APFD)

Definition (Approximate Predictive Functional Dependency (APFD))

Given a TF-view $w = TFv(mtr, \alpha, m, [1, k])$ of an mt-relation mtr with schema $Z\overline{U}^1\overline{U}^2..\overline{U}^k\dot{B} \cup \{\overline{VT}^1, \overline{VT}^2, .., \overline{VT}^k, \dot{VT}\}$, w fulfills the APFD

$$\boldsymbol{S\overline{P}^h\overline{Q}^i ...\overline{R}^j \xrightarrow[\alpha,m]{\varepsilon} \dot{Y}}$$

(written as $w \models^{E}_{\alpha,m} S\overline{P}^h\overline{Q}^i \ ...\overline{R}^j \xrightarrow{\varepsilon} \dot{Y})$, where $\varepsilon = \langle \varepsilon_g, \varepsilon_h, \varepsilon_j \rangle$ and $S \subseteq Z, \overline{P}^h \subseteq \overline{U}^h, \overline{Q}^i \subseteq \overline{U}^i, \overline{R}^j \subseteq \overline{U}^j, \ \dot{Y} \subseteq \dot{U}$, if a relation $s \subseteq w$ exists such that $s \models^{E}_{\alpha,m} S\overline{P}^h\overline{Q}^i \ ...\overline{R}^j \rightarrow \dot{Y}$ with $g_3 \leq \varepsilon_g \wedge h_3 \leq \varepsilon_h \wedge j_3 \leq \varepsilon_h$.

$\varepsilon_g, \varepsilon_h, \varepsilon_j$ are the maximum acceptable errors defined by the user for g_3, h_3, j_3 respectively.

Table of Contents

The (data) complexity of deriving an APFD

To discuss the complexity of checking an APFD, it is enough to consider a relation having:

- a single attribute (A) representing the antecedent;
- the predicted attribute ($\dot{B}$);
- a single attribute (Z) representing the entity attribute.

The domain of all attributes is $\mathcal{N}$ or a subset of it (the predicted values for $\dot{B}$ will be either 0 or 1, to represent boolean values).

Thus, we will consider a relation w with schema

$$\boldsymbol{W(A, \dot{B}, Z)}$$

The (data) complexity of deriving an APFD

Given a relation $w \subset \mathbb{N}^3$, a natural number $0 \leq \mathbf{k} < |w|$, and a natural number $0 \leq \mathbf{h} < |\pi_Z(w)|$ determine whether or not w admits a *conflict resolution* of order $(\mathbf{k}, \mathbf{h})$.

k represents the threshold G_3.

h represents the threshold H_3.

We prove that the problem of verifying any APFD even only considering H_3 is **NP-Hard**. (Proof by reduction from an already known problem [1].)

[1] Christos H. Papadimitriou and Mihalis Yannakakis. Optimization, approximation, and complexity classes. Journal of Computer and System Sciences, 1991.

The (data) complexity of deriving an APFD

We reduced the problem in hand to a general $3SAT$ problem, showing that checking an APFD considering all the three thresholds belongs to the class NP.
An instance of $3SAT$ problem is a logical formula formed by a conjunction of disjunctive clauses, where each clause has exactly 3 literals.

$$(X_1 \vee X_2 \vee X_3) \wedge (X_4 \vee X_5 \vee X_6) \wedge (X_7 \vee X_8 \vee X_9)$$

The (data) complexity of deriving an APFD

After proving that verifying any APFD even only considering H_3 is NP-Hard, we propose a deterministic algorithm that could stop the analysis of a relation, as soon as it verifies that the relation cannot satisfy the given APFD.

General idea: searching for a solution considering one tuple at a time, until it is possible to generate a solution, which satisfies the selected thresholds.

Table of Contents

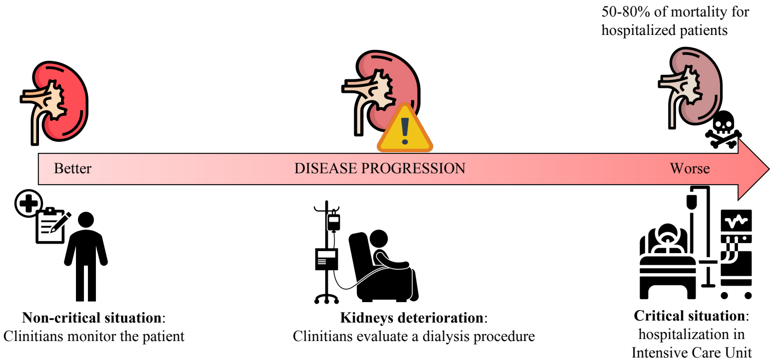
Experimental evaluation: the application domain

Intensive care unit

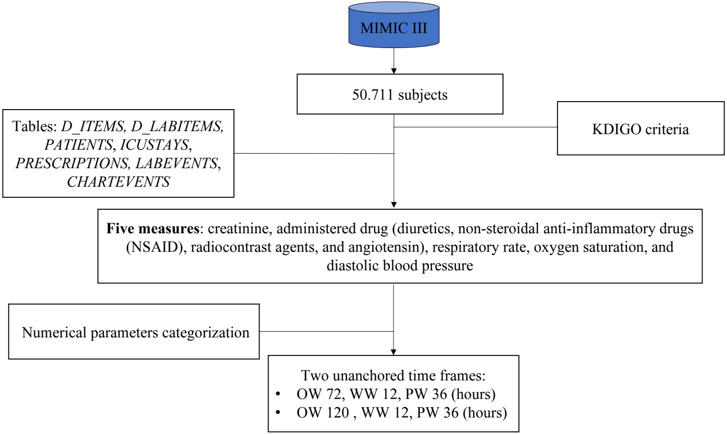
- Physicians have the access to a **large quantity of data** for each patient, derived from the continuous monitoring.
- **Timing** is a fundamental part: anticipation of the illness onset, worsening of clinical condition or the diagnosis moment.
- It could be difficult to **identify knowledge** for clinical decisions: data mining techniques are useful to identify the most significant information.

Experimental evaluation: the application domain

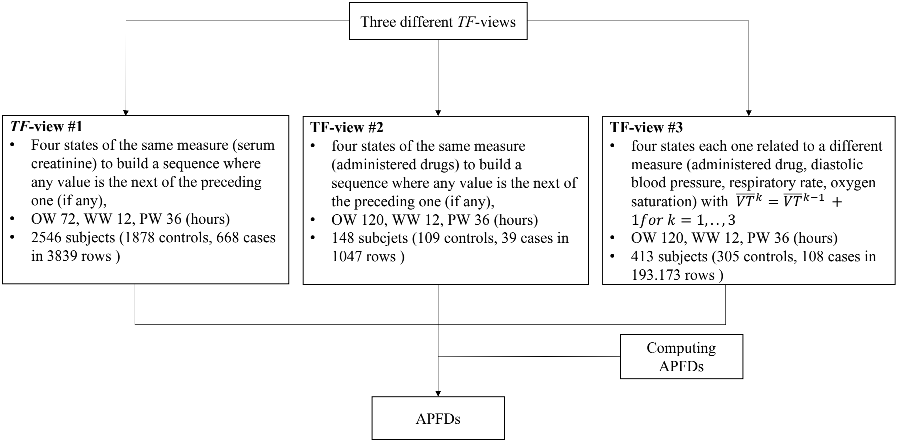
AKI is a syndrome characterized by sudden kidney failure (high values of creatinine and low urine output) with a rapid progression.



Experimental evaluation: dataset and preprocessing



Experimental evaluation: dataset and preprocessing



We report some of the APFDs obtained through the algorithm, with the corresponding error thresholds.

APFD	ε_g	ε_h	ε_j	TF-view
$\overline{Creat}^1, \overline{Creat}^3 \rightarrow \dot{AKI}$	27.45%	27%	50%	#1
$\overline{Creat}^1, \overline{Creat}^4 \rightarrow \dot{AKI}$	27.45%	27%	50%	#1
$\overline{Drug}^1, \overline{Drug}^2, \overline{Drug}^4 \rightarrow \dot{AKI}$	21%	30%	50%	#2
$\overline{Drug}^1, \overline{Drug}^2, \overline{Drug}^4 \rightarrow \dot{AKI}$	21%	30%	80%	#2
$\overline{Drug}^1, \overline{Drug}^2, \overline{Drug}^3 \rightarrow \dot{AKI}$	21%	30%	80%	#2
$\overline{Drug}^1, \overline{Drug}^3, \overline{Drug}^4 \rightarrow \dot{AKI}$	21%	30%	80%	#2
$\overline{Drug}^1, \overline{RespRate}^3 \rightarrow \dot{AKI}$	10%	51%	75%	#3
$\overline{RespRate}^3 \rightarrow \dot{AKI}$	30%	75%	75%	#3
$\overline{Drug}^1 \rightarrow \dot{AKI}$	30%	75%	75%	#3
$\overline{Spo_2}^4 \rightarrow \dot{AKI}$	30%	75%	75%	#3

Table of Contents

Conclusions

We proposed a methodology for deriving a new kind of approximate temporal functional dependencies, called **Approximate Predictive Functional Dependencies**.

- A formal 3-window model to derive the APFDs;
- The computational aspects of deriving an APFD;
- The application to real clinical data, specifically to MIMIC III dataset.

Thank you for your attention!

Extra slides

The applicability of our framework

We can apply the entire framework in every domain where the prediction task could be an interesting task.

The usefulness of the 3-widow model is tied to two aspects:

- the final goal related to the problem in hand;
- the nature of the predicted event.

Why a 3-window model

The nature of the predicted event.

The waiting window is used to anticipate an action in order to prevent a future event. It is important to underline that not every type of events could be prevented.

Example in medicine: Diabetes diagnosis, we cannot prevent this diagnosis, because it's a fact that simply happened at a certain point, and we cannot avoid it.

AKI, Sepsis, Covid-19 are diseases that imply a possible deterioration or improvement of the patient status. So in this case, the waiting window could be use to anticipate as soon as possible the diagnosis, preventing the deterioration.

Why a 3-window model

The final goal related to the problem in hand.

Using again the diabetes diagnosis. Suppose to have a database that records EHR from a childhood diabetes center.

A way to use our model could be consider the final goal to study all the different temporal events such as specialist visit, hospitalization in the emergency department, in order to anticipate the start of the cure of these patients.

In this case it is not possible to prevent an event (diabetes) that is unavoidable, but we can the anticipate the moment of the diagnosis, the start of the treatment in order to alleviate the long term side-effects.

Predictive Functional Dependency (PFD)

A PFD holds on an mt-relation *mtr* with schema *MTR* in a timeframe *TF* with modality *m*, *with a restricted or extended range semantics* (denoted as $mtr \models^{R}_{\alpha,m}$ *or* $mtr \models^{E}_{\alpha,m} S\overline{P}^h\overline{Q}^i \ldots \overline{R}^j \rightarrow \dot{Y}$) iff:

$$\forall t, t' \in mtr((t[S\overline{P}^h\overline{Q}^i \ldots \overline{R}^j] = t'[S\overline{P}^h\overline{Q}^i .. \overline{R}^j] \wedge \Theta(t, \alpha, m, [h, j]) \wedge \Theta(t', \alpha, m, [h, j])) \rightarrow t[\dot{Y}] = t'[\dot{Y}])$$

or

$$\forall t, t' \in mtr((t[S\overline{P}^h\overline{Q}^i \ldots \overline{R}^j] = t'[S\overline{P}^h\overline{Q}^i .. \overline{R}^j] \wedge \Theta(t, \alpha, m, [1, k]) \wedge \Theta(t', \alpha, m, [1, k])) \rightarrow t[\dot{Y}] = t'[\dot{Y}])$$

Computing APFDs

We compute all the APFDs, adopting a tractable sub-optimal solution and considering the three errors, g_3, h_3, j_3.
Given a KSPE instance w and the predicted attribute $\dot{B}$, our approach is mainly based on the following steps:

- Derive s by TANE, such that $g_3 \leq \varepsilon_g$;
- Check on s that $h_3 \leq \varepsilon_h$;
- If the previous check is fine, check $j_3 \leq \varepsilon_j$.