# Exploration of patterns predicting renal damage in patients with diabetes type II using a visual temporal analysis laboratory

Denis Klimov, Alexander Shknevsky, Yuval Shahar

# ABSTRACT

**Objective** To analyze the longitudinal data of multiple patients and to discover new temporal knowledge, we designed and developed the Visual Temporal Analysis Laboratory (ViTA-Lab). In this study, we demonstrate several of the capabilities of the ViTA-Lab framework through the exploration of renal-damage risk factors in patients with diabetes type II.

**Materials and methods** The ViTA-Lab framework combines data-driven temporal data mining techniques, with interactive, query-driven, visual analytical capabilities, to support, in an integrated fashion, an iterative investigation of timeoriented clinical data and of patterns discovered in them. Patterns discovered through the data mining mode can be explored visually, and vice versa. Both analysis modes are supported by a rich underlying ontology of clinical concepts, their relations, and their temporal properties. The knowledge enables us to apply a temporal-abstraction pre-processing phase that abstracts in a context-sensitive manner raw time-stamped data into interval-based clinically meaningful interpretations, increasing the results' significance. We demonstrate our approach through the exploration of risk factors associated with future renal damage (micro-albuminuria and macro-albuminuria) and their relationship to the hemoglobin A1C (HbA1C) and creatinine level concepts, in the longitudinal records of 22 000 patients with diabetes type II followed for up to 5 years.

**Results** The iterative ViTA-Lab analysis process was highly feasible. Higher ranges of either normal albuminuria or normal creatinine values and their combination were shown to be significantly associated with future micro-albuminuria and macro-albuminuria. The risk increased given high HbA1C levels for women in the lower range of normal albuminuria, and for men in the higher range of albuminuria.

**Conclusions** The ViTA-Lab framework can potentially serve as a virtual laboratory for investigations of large masses of longitudinal clinical databases, for discovery of new knowledge through interactive exploration, clustering, classification, and prediction.

.....

Key words: Visual Analytics; data analysis; temporal data mining; knowledge discovery; temporal abstraction; ontologies

#### **INTRODUCTION**

Effective analysis of time-oriented multivariate clinical data, with the objective of investigating processes and predicting their course, as is important in the case of diabetes, requires the combined use of multiple approaches, including mining the longitudinal clinical data to automatically discover within it meaningful patterns, and exploring it interactively in a userdriven fashion. In this paper, we introduce a general framework that combines both approaches, which we refer to as the Visual Temporal Analysis Laboratory (ViTA-Lab). We will demonstrate our new framework by investigating the process of the development of renal dysfunction in patients with diabetes (manifested as micro-albuminuria and macro-albuminuria), with the objective of predicting its course. Diabetes affects more than 170 million people worldwide. About one-third of those affected will eventually have progressive deterioration of renal function.<sup>1</sup> The first sign of renal dysfunction in patients with diabetes is micro-albuminuria (urinary albumin excretion), affecting 20–40% of patients 10–15 years after the onset of diabetes type II, and 30–60% of patients with type I diabetes.<sup>2</sup> Macro-albuminuria (overt nephropathy) is present in 20–40% of patients 15–20 years after the diabetes onset. Early detection of nephropathy, or of patients at high risk for it, is important for improving the outcome of diabetes.

The ViTA-Lab framework combines computational datadriven interval-based temporal data mining (TDM) techniques, with interactive, user-driven visual analytical capabilities for a knowledge-based investigation of time-oriented clinical data.

Correspondence to Denis Klimov, Medical Informatics Research Center, Department of Information Systems Engineering, Faculty of Engineering Sciences, Ben-Gurion University of the Negev, PO Box 653, Beer-Sheva 84105, Israel; klimov@bgu.ac.il

© The Author 2014. Published by Oxford University Press on behalf of the American Medical Informatics Association.

All rights reserved. For Permissions, please email: journals.permissions@oup.com

For affiliation see end of article.





RECEIVED 1 May 2014 REVISED 15 September 2014

ACCEPTED 20 September 2014

PUBLISHED ONLINE FIRST 28 October 2014

The concurrent application of both ends of the analytical spectrum (ie, goal-driven visual analysis and data-driven TDM) supports an iterative process for the discovery and exploration of new, meaningful temporal patterns, in longitudinal, multivariate clinical data, integrating the best of each of the two worlds:

- Interactive visual-exploration systems provide users with an overview of the data, enabling them to explore the visualized data to answer user-initiated queries; they are user friendly and focus on concepts that are highly promising for meaningful knowledge discovery. However, the user must know what to look for, and which questions to ask. If a query about a key association has not been asked, a potentially important pattern might be missed. Thus, query-driven methods are *precise*, in the sense of producing mostly significant answers, but are often *incomplete*.
- Pure computational data mining (DM) methods (in particular, TDM methods) are automated, computationally valid, and complete (ie, they discover all temporal patterns that can be found in the data); but most of them are not interactive, are intended only for a 'super-user' with a significant experience, and do not allow an effective exploration of the (typically too numerous) computed output, much of which is irrelevant. As a result, some significant insights might be missed. Thus, data-driven methods are *complete*, but their *precision* is low, in the sense that most discovered patterns are of low significance.

Therefore, the ViTA-Lab framework combines both types of analysis; that is, query driven and data driven. The combination of visual and analytical capabilities of data analysis has been referred to as visual analytics (VA).<sup>3,4</sup> Thus, one might refer to our approach as a special, time-oriented type of a VA system.

To significantly enhance the capabilities of a VA system, we propose to first pre-process the input raw, time-stamped data, such as hemoglobin values at particular times, using relevant. context-sensitive, domain-specific knowledge, to produce a set of clinically meaningful summarizations and interpretations, such as a 3-week period of moderate anemia in the context of a 17-year old man, known as temporal abstractions (TAs). In general, TA is the aggregation of a time series into a succinct, symbolic, typically interval-based representation, suitable for a human decision-maker, or for DM. To abstract the raw data, one can use prior domain-specific clinical knowledge, or can proceed using various computational methods.<sup>5</sup> In our studies, we usually apply the knowledge-based temporal-abstraction (KBTA) method,<sup>6</sup> which uses domain-specific temporalabstraction knowledge to compute TAs. The KBTA method is applied by an enhanced version of the IDAN<sup>7</sup> temporal mediator, which mediates between decision-support applications and a time-oriented clinical database. When the knowledge necessary to specify abstractions is not available, we apply static generalized discretization methods, such as equal-width discretization and equal-frequency discretization, or more complex methods geared specifically for temporal

discretization, such as Symbolic Aggregate approXimation (SAX)<sup>8</sup> and Persist.<sup>9</sup>

The KBTA method and the generic computational discretization methods underlying our work are domain independent, since knowledge-based interpretation (when domain-specific knowledge is provided) and generic abstraction methods can be applied in any medical (or non-medical) domain. Examples of previous applications of this highly general ontology include domains such as oncology,<sup>10</sup> diabetes,<sup>11,12</sup> monitoring children's growth,<sup>13</sup> and prenatal monitoring,<sup>14</sup> and non-medical domains, such as information security.<sup>15,16</sup> Thus, the ViTA-Lab framework is domain independent and can be used to analyze any type of longitudinal data.

To facilitate the introduction of the semantics and functionality of the ViTA-Lab framework, we will present in this study, in detail, a case, in which, given the records of 22 000 anonymous patients with type II diabetes followed sporadically over 5 years at our university's academic medical center, we explored the association between several potential risk factors and future renal injury.

#### BACKGROUND AND SIGNIFICANCE

#### Visual analytics in the medical domain

Previously, visual exploration systems in medical domains focused mostly on the visualization of raw longitudinal data for individual<sup>17,18</sup> or multiple<sup>19,20</sup> patient records, as reviewed by Chittaro<sup>21</sup> and recently by Rind *et al.*<sup>22</sup> Information-visualization methods have often focused on the development of innovative interfaces, graphical metaphors, and exploration capabilities, rather than on the discovery of actual new knowledge, as noted by Aigner *et al.*<sup>23</sup> Furthermore, most VA studies focus on representation and analysis of *static* data, such as text, networks, and sparse quantitative data, while the *temporal* aspect of VA is less studied, due to its complexity, as noted by Aigner<sup>24</sup> in the context of medical applications.

Recent visual exploration systems include additional capabilities for sophisticated interactive exploration of multiplepatients' data.<sup>25,26,27</sup> However, most VA systems focus on *raw* data, such as a time series of laboratory test results, rather than on its *interpretations*. They include neither an underlying domain-specific knowledge base that formally represents the explored concepts and the relationships among them, nor any computational mechanisms that could capitalize on such knowledge to produce derived concepts (interpretations) for visualization and exploration; and they do not support an *iterative* exploration of discovered knowledge, for example, apply a pattern, which was just discovered in a group of patients, to another patient population.

Several recent studies focus on the analysis of temporal event sequences; examples include EventFlow<sup>28</sup> for individual patients, LifeFlow,<sup>29</sup> OutFlow,<sup>30,31</sup> CareFlow<sup>32</sup> for multiple patients, and an advanced framework, Frequence,<sup>33</sup> which also provides capabilities for frequent pattern mining. Although these systems provide advanced visual capabilities for aggregation of multiple sequences, such as care processes, their main focus is on exploration, while the ViTA-Lab framework

focuses on the discovery of new knowledge, with further reuse of the discovered knowledge to perform further analysis on the patients' data. Moreover, these systems mostly support sequences ordered only by the 'after' temporal relation, that is, event B occurs after event A. However, real interval-based clinical data can be much more complex; for example, treatment by two medications might proceed in parallel, both periods overlapping a period of exacerbation of the patient's clinical state. Thus, more sophisticated DM algorithms that support relations such as 'overlaps', 'during', and all of Allen's<sup>34</sup> other temporal relations are required.

# Assessing the risk for future renal damage in patients with diabetes

Several studies have focused on prediction of abnormal levels of albuminuria, using DM computational methods.<sup>35,36,37,38,39,40</sup> However, they usually do not exploit domain knowledge: the DM algorithms are applied to raw, time-stamped data; they lack an iterative application: discovered knowledge cannot be reused for further analysis; and they lack visual exploration capabilities: the results of the mining process cannot be easily explored by a dedicated interface.

#### METHODS: THE VITA-LAB FRAMEWORK

# Application of the ViTA-Lab framework in different clinical domains

To apply the ViTA-Lab analytical framework in a new domain, we need to perform (once) two steps:

- 1. Knowledge acquisition. The knowledge provided by the domain expert is graphically specified by a dedicated knowledge-acquisition tool, called Gesher.<sup>41</sup> Although one of the major sources for the definition of medical concepts in our framework is textual clinical guidelines, we are not applying an automated process for extraction of temporal relations from text, as for example proposed by Tao et al<sup>42</sup> or by Kaiser and Miksch.<sup>43</sup> since several of the properties required by the KBTA ontology are usually not represented in the text (eg, the validity-persistence time of a laboratory measurement). The Gesher tool supports the specification of knowledge concepts by multiple standardized vocabularies, including UMLS, LOINC, Rx-NORM, and CPT. The vocabularies' codes serve the user in the next step, which we refer to as the *mapping step*—mapping the knowledge base to the local data base (DB).
- 2. Mapping the knowledge to the local data. Although the ViTA-Lab framework suggests a generic format for data records, in some cases we need to use an additional mapping component to create a link between the records in the local DB and knowledge concepts to analyze the data by the ViTA-Lab tools. The link between the necessary records in the local DB and knowledge concepts is established by using the vocabularies' codes as specified in the raw-data leaves of the knowledge base.<sup>44</sup> Some mapping tools such as KDOM<sup>45</sup> support the generic data models, for example vMR, for the knowledge-data mapping.

#### The main interfaces of the ViTA-Lab framework

The ViTA-Lab framework includes three main visual interfaces, as shown in figure 1.

- 1. The main visualization and exploration interface (denoted by '1' in figure 1) provides an interactive overview of the raw longitudinal concepts and of the TAs for individual and multiple patients. The left panel of the interface includes a knowledge-based browser showing the domain's ontology. in this case in the diabetes domain; and a graphical widget for selection of the patients by providing an ID, in the case of individual patients, or by providing the name of a group of patients previously selected according to a set of a demographic and knowledge constraints.<sup>46</sup> Clicking on a concept in the browser's ontology tree for the selected group of patients shows either the values of a raw time-stamped concept, or the distribution, for each time granule of a derived TA. For instance, the top panel of the visualization interface shows a scatter diagram over time (the horizontal axis) of the blood creatinine values (the vertical axis) for a group of male patients who were selected (in another interface, not shown) for having their first albuminuria measurement being 'Normal' (denoted as 'male FirstNormo'). Each point represents a specific measurement of one patient. In addition, the maximal, minimal, and mean values among the patients are computed according to the temporal granularity (eq. month) in which the user currently chooses to explore the data: the granularity can be interactively modified. The second panel from the top shows the visualization of the alucose-state TAs for the same group of patients. TAs for multiple patients are displayed as a distribution of the values of the abstract concept of the patients at the current temporal granularity. For the visualization, we use a modified version of the bar chart visualization technique. The modification includes providing separate [0...100%] scales for each of the TAs (eg, for the levels of the glucosestate abstraction), which is useful for discovering trends in the distribution. Several visualization operators enable the user to explore the data of raw and abstract concepts at any temporal granularity, including a relative time line which refers, as its zero point, to some key event such as intervention. The distributions of the abstractions are computed on-the-fly during the exploration. The full description of the VISITORS framework and its displays is outside of the scope of the current paper, and can be found elsewhere.47
- 2. The Temporal Association Chart (TAC) (denoted by '2') enables the user to visually explore probabilistic temporal associations among the distributions of multiple different raw and, in particular, abstract concepts at different times. Usually a TAC includes two or more different concepts explored using the main visualization interface; thus, the input for the TAC is the group of patients and list of concepts (elements of the TAC) selected within the same or a different time window panel. The distributions of values for each concept are computed within the selected time

**Figure 1**: A typical set of screens from the Visual Temporal Analysis Laboratory (ViTA-Lab) for analysis and exploration of the time-oriented data of multiple patients. (1) The main visualization and exploration interface provides an interactive overview of the raw longitudinal concepts, and of the distribution of the derived temporal abstractions for multiple patients, at any temporal granularity. (2) The Temporal Association Chart (TAC) enables the user to visually discover probabilistic temporal associations among the distributions of multiple different concepts at different times. (3) The Patterns Explorer displays frequent temporal patterns in the data and supports their exploration.



period. The corresponding data values for each patient between two consecutive elements are linked. Several links involving the same pair of values for multiple patients (eg, the probability of having the Moderate Anemia value of the hemoglobin-state concept, given the low value of the white-blood-cell-state concept) are aggregated into a temporal association rule that denotes the probability of having the value of the second concept, given the value of the first concept (ie, confidence), and the overall frequency of that combination (ie. *support*). Each rule represents a set of patients who have had this particular combination of values for the two concepts, at the same time or at different times, depending on the time period(s) in which the user is interested. (The exploration example in the Results section demonstrates several of the probabilistic and statistical values computed for each link). The order between the concepts appearing in a TAC can be changed to analyze the associations between the different concepts of the TAC. For user convenience, the left panel shows the color legend for each

concept in TAC. The full description of the TAC functionality and analysis capabilities is outside the scope of this paper, and can be found elsewhere.<sup>48</sup> The semantics of the width and color of the association links are explained in the caption to figure 4. An enhanced version of TAC enables the user to explore the associations between two concepts at arbitrary temporal granularities.<sup>49</sup> Examples of using enhanced TACs include exploring adverse drugs events within a pharmacovigilance project, or assessing the outcomes of the application of a particular clinical guideline. The enhanced version also provides a method to display various statistical measures of the specific association over time, such as whether the association strength increases, or whether the number of patients involved decreases over time.

 To support the exploration of numerous temporal patterns that can be discovered by the data-driven computational process, we have designed and developed a dedicated interface, called the Patterns Explorer (denoted by '3'). Its underlying semantics are based on a version of the KarmaLego algorithm for the discovery of frequent temporal patterns.<sup>50,51</sup> The discovered output temporal patterns are represented in corresponding panels (see figure 2), each pattern being displayed in a specific panel. The color of the same type of component in a pattern is the same across all patterns and is represented in the legend on the left side in the Patterns Explorer. To focus on most relevant patterns from a large number of discovered patterns, we use filtering mechanisms based on regular expressions; for example, find all frequent patterns that include a specific pair of interval-based concepts with a specific temporal relation between them.<sup>34</sup> The basic symbol in the regular expression can include a concept, for example, HbA1C-state, or a concept and value, for example, HbA1C-state = 'High', one of Allen's seven basic temporal relations to specify the interrelationship between pattern components, and a wildcard to specify exactly one symbol ('?') or any number of symbols ('\*') in a pattern. In addition, we provide two additional temporal constraints over components of patterns, either 'start by' or 'end with'. For example, in the use case described below, we looked for patterns ending in microalbuminuria or macro-albuminuria.

The ViTA-Lab framework enables users to explore the longitudinal data of multiple patients, using the Visual Information Mantra<sup>51</sup> and the enhanced Visual Analytics Mantra.<sup>3</sup> Typically, the first step in an analytical process is the overview of the patients' data to better understand overall trends, allowing users to zoom into or zoom out of specific time periods, interesting concepts and group of patients; thus, in that step, users usually use the main visualization interface.

Then, users can apply the TAC module to selected concepts and patients to better understand the relationship between the distributions of various concepts over time, or they can apply the pure DM functions to discover frequent patterns within the overall population or a specific sub-group of patients. Note that the DM step can be applied as the first step, but that is often ineffective, due to the very large number of discovered patterns (applied on all of the population and for all possible concepts).

In the TAC interface, the user is able to filter the various associations discovered between pairs of concept-value distributions, by statistical values (eg, by requiring a minimal level of support), by selection of a specific value in the concepts displayed in a TAC. This is, in fact, how we have explored the relationship between various levels of normal albuminuria values in the first year, and micro-albuminuria in the fifth year of follow up of the same group of patients with diabetes. Furthermore, the user is also able to apply a DM process to the selected patients and/or concepts to understand what additional patterns characterize the selected patients.

To sum up, we are emphasizing in our methodology the importance of an iterative process of data analysis, in which both analysis types, [visual] query driven and data driven, are integrated into one comprehensive virtual laboratory for time-oriented data analysis and exploration—the ViTA-Lab framework.

We now demonstrate the working process and the capabilities of the ViTA-Lab framework through a specific case study, exploration and prediction of renal dysfunction in patients with diabetes.

**Figure 2**: A graphical representation of a pattern in the Pattern Explorer whose informal definition is 'for 6% (a Vertical Support (VS) of 0.06) of the male patients, an episode of Normal Creatinine-State overlaps an episode of Normal HbA1C-State, which is followed by an episode of a Moderately-High value of HbA1C State'. Each component of a pattern (a pair of concept and value) is represented by a horizontal line, the lines being ordered in the panel according to the start time point of each component (the earliest interval is the first from the left), maintaining, in a proportional fashion, the mean duration (across all pattern instances) of each component and of the gaps among components. Thus, the user can recognize visually, in a very simple manner, the meaning of a discovered temporal pattern, that is, which components exist in pattern, and what temporal relations such as before, after, or overlaps, hold between them. This pattern is valid for 6% of the patients, as shown by the level of VS (ie, across all patients) in the part of the panel denoted by '1'. The temporal relations between the pattern's components are shown in the part of the panel denoted by '2'. Selection of a pair of components and a relation from the list will highlight the selected components. In addition, the mean duration of each component, and the mean gap between components, are computed and shown (in the part denoted by '3'). HbA1C, hemoglobin A1C.



## MATERIALS

In the current study, we explored the data of a group of 22 160 anonymous patients with diabetes (the total number of raw data records is close to 12 million) from our local academic medical center, who had been followed (albeit sporadically) for at least 5 years. Our focus here is on the investigation of factors associated with changes in renal function (mostly focusing on the level of albuminuria, or secretion of protein in the urine), exploring its predictive risk factors, and whether there are differences in the behavior of that concept over time between male and female patients.

A knowledge base, including the ranges of albuminuria for women and men (see figure 3), HbA1C (see table 1), and creatinine levels (see table 2) was specified by a medical domain expert in the Gesher knowledge acquisition tool.

However, during the analysis we found that the expert's suggested value ranges for creatinine-state 'Normal' level (<1.2 mg/dL for men and <1.0 mg/dL for women) did not lead to any significant associations or predictive patterns. Thus, we created by discretization (by using the TD4C classification-driven temporal-discretization method<sup>53</sup>) three sub-ranges of the 'Normal' range. The ranges for creatinine state for male and female patients are represented in table 2. For similar reasons, we have added the Normo-Low and Normo-High sub-ranges within the normal (0–30  $\mu$ g albumin/mg creatinine) albuminuria values.

#### RESULTS

In this section, we present the results of performing a visual analysis of renal dysfunction in patients with diabetes using the ViTA-Lab framework.

# A general overview of the albuminuria-state concept for male and female patients over a period of 5 years

We will start our iterative exploration of the renal-injury complication in the diabetes domain by using the query-driven visual interface.

Figure 4 presents an overview of the albuminuria-state value distribution over time (from 2004 to 2008) for a group of 11 105 male patients (cumulatively, ie, not all of the patients have data for each year), using a TAC. In this interface we can see an overview (the first step in the visual information mantra) and an analysis of temporal associations (the first step of the VA mantra). Note that the values shown in the visualization for each year need not necessarily exist *throughout* the year; rather, they represent, or *characterize*, the overall abstractions that can be derived for each of the relevant patients during that year. The semantics of the width and color of the association links are explained in the caption to figure 4.

# Table 1: The HbA1C-state abstract concept levels and the corresponding values of the HbA1C raw concept

HbA1C values	HbA1C-state levels
<7	Normal
7–9	Moderately-High
9–10.5	High
>10.5	Very-High

HbA1C, hemoglobin A1C.

**Figure 3**: The OR-based (disjunction) derivation tree for the albuminuria-state abstraction, for male and female patients, displayed as a two-level abstraction function, in which the top level uses the definitions of the bottom level. ACR = albumin/creatinine ratio ( $\mu$ g albumin/mg creatinine). Note that the derivation tree allows for computation of the Albuminuria values in two different ways, depending on available input in the local database.

			Mal	e						Ferr	ale		
Albumi (ACR)_Sta	nuria te Levels		Albu (A24h)_	minuria State Levels	Albuminuria_State Levels		Albumin (ACR)_Stat	nuria te Levels		Albu (A24h)_	minuria State Levels		Albuminuria_State Levels
Normo	-Low	OR	Nor	$mo-Low \Rightarrow$	Normo-Low		Normo-	Low	OR	Nor	no-Low	⇒	Normo-Low
Normo	High	OR	Nor	no-High $\Rightarrow$	Normo-High		Normo-	High	OR	Nor	no-High	⇒	Normo-High
Mic	ro	OR	N	1icro ⇒	Micro		Mici	o	OR	N	1icro	⇒	Micro
Mac	ro	OR	N	íacro ⇒	Macro		Mac	0	OR	Ν	facro	⇒	Macro
							<						
Microalbumin	Albu	min	uria	Microalbumin	Albuminuria	Mie	croalbumin_	Albu	min	uria	Microalbun	nin	Albuminuria
creat values	(ACR)_S	tate	Levels	U24h values	(A24h)_State Levels	CI	reat values	(ACR)_S	state	e Levels	U24h valu	es	(A24h)_State Levels
0-13	Norn	no-L	ow	0-13	Normo-Low		0-15	Normo-Low		0-15		Normo-Low	
13-30	Norn	no-H	ligh	13-30	Normo-High		15-30	Normo-High		ligh 15-30			Normo-High
30-300	M	licro	)	30-300	Micro		30-300	Micro		Micro 3			Micro
>300	M	acro	)	>300	Macro		>300	M	lacro	)	>300		Macro

Table 2: The creatinine-state abstract concept levels for male (left) and female (right) patients, and the corresponding values of the creatinine raw concept

Creatinine values	Creatinine-state male levels	Creatinine values	Creatinine-state female levels
<0.8	Low	<0.56	Low
0.8–0.9	Normo-Low	0.56–0.7	Normo-Low
0.9–1.1	Normo-High	0.7–0.8	Normo-High
1.1–1.4	Normo-Severe	0.8–1.1	Normo-Severe
1.4–2.0	High	1.1–2.0	High
2.0–4.0	Very-High	2.0–4.0	Very-High
>4.0	Severe	>4.0	Severe

Three main 'temporal clusters' or 'temporal pathways' can be identified visually in figure 4.

- From 2004 to 2008 the portion of patients with 'Normo-Low' value of the albuminuria-state concept is going up from 31% in 2004 to 53% in 2008. Although the overall data do not necessarily represent the same group of patients at each point in time (originally, the clinical measures were sampled for somewhat different patients within the overall patient group, each year), the TACs display only patients who have had values throughout the whole period. The overall trend seems to be towards an increase in the portion of patients with lower albuminuria levels (denoted by '3'), perhaps signifying improved management of patients with diabetes.
- 2. Similarly, we can identify each year, over the 5 years, a group of about 700 patients with a 'Micro' value of the albuminuria-state concept (denoted by '4'). (Note that these patients are not necessarily the same ones every year; the confidence measure displays the probability of the same patient having the same value from year i to year i+1).
- 3. Finally, we focus on a group of approximately 300 patients each year who have the 'Macro' values of the albuminuriastate concept (denoted by '5'). Although the typical support value for staying in that state from one year to the next is low (less than 3% of the patients had macro-albuminuria in year i and in year i+1), the *confidence* (ie, the conditional probability) of maintaining that state is around 50%; that is, once a patient has had a 'Macro' level of the albuminuriastate concept, their probability of either staying at that level, or of improving with respect to that measure, is around 50%. Not surprisingly, given this high persistence, the lift (relative risk) measure from one year to the next for having a 'Macro' value in the next year, given a 'Macro' value of the albuminuria-state concept in the previous year, is very high, denoting a high relative risk (5.2-6.3 times, vs the expected neutral risk of 1).
- 4. Similar pathways could be displayed for female patients (not shown here due to lack of space).

However, the view using the complete absolute time line (ie, from 2004 to 2008) cannot answer the question regarding the factors predicting a worsening of the renal function, since only a small group of patients have had their data consistently measured across all of these years. To examine *all* of the sufficiently frequent temporal patterns, we will now apply a purely data-driven, computational temporal DM method.

# Differences in progression of albuminuria states between male and female patients

The ViTA-Lab DM engine was applied separately for men (2356 patients) and women (2540 patients) who have had albuminuria, creatinine, and HbA1C raw data values measured over up to 5 years of therapy.

The resulting discovered frequent (above a given threshold) patterns, consisting of any two consecutive measurements with a maximal time gap between them of not more than 1 year, over all of the 5 years, are represented in figure 5.

As we can see, in addition to mostly similar values of support between male and female patients, the time gaps between the different albuminuria values for the same patterns are also mostly similar. For example, the mean time gap denoting a worsening from a 'Normo-Low' to a 'Micro' value of albuminuria-state is 10.4 months for men (pattern denoted by '1a') and 10.24 months for women (pattern denoted by '2a'). Similar gap periods can also be noted in the case of the rest of the albuminuria pair-of-value patterns.

The temporal-gap similarity might occur randomly, but might also represent a real similarity between men and women with respect to the patterns denoting a worsening of the albuminuria levels. Although an additional, more detailed analysis is required, it is clear that within a few minutes one can obtain preliminary intuitions and a hypothesis to test by further investigations.

We then examined the possibility of using the DM engine for predicting the clinically interesting micro/macro-albuminuria state(s), which are evidence for renal damage, in the fifth year of follow-up, based on temporal patterns that are discovered from state abstractions derived from the raw values of

Figure 4: Temporal associations of abstracted albuminuria levels over 5 years, for 11 105 male patients. The distribution of the four albuminuria levels (the color-based legend of values appears on the left side), that is, the relative proportion of patients who have had each albuminuria-state in each year of the follow-up, is represented within the specific time interval (represented on the bottom). In this case, we chose an annual (1 year) temporal granularity view, that is, we selected the start and end times of each frame so as to represent the absolute time periods of five consecutive years. Thus, the annual albuminuria-state for 1537 (29.59%) of the patients during 2004 was summarized as 'Micro' (see the yellow tooltip denoted by '1'). The red links between two values of different adjacent distributions represent a group of patients who have had both of the values, and thus represent an association over time between these values. The deeper the shade (the hue) of the link's color, the higher the level of confidence in the relationship. That is, a darker shade of red indicates a higher confidence level in the future (right) value, given the current (left) value. The width of the link corresponds to the level of support, that is, the number of subjects having this particular value combination for these two time windows; broader links represent an association with higher support. By hovering with the mouse over an edge, the user will see additional statistical information, such as the lift (relative risk) and the statistical proportion test and its significance (see vellow tooltip denoted by '2'). In this case, the tooltip represents the relation between the 'Normo-Low' values of the albuminuria-state concept between 2004 and 2005 years: 6.12% of the patients in the relevant patient group had this particular combination of values (ie, support = 0.612), and 42.26% of the patients with 'Normo-Low' albuminuria-state values during the 2004 year also had 'Normo-Low' albuminuria-state values during 2005 (ie, confidence = 0.4226). This temporal association was valid for 680 patients, with a lift (a relative risk) of 1.48 (ie, patients are more likely than expected just by the prior probability, to stay within the 'Normo-Low' albuminuria range from one year to the next). A proportion test that compares the confidence of the association (42.26%) with the prior probability of the albuminuria-state having the value 'Normo-Low' in the year 2005 (28.58%), using the actual patient numbers, was significant with p < 0.05 ('True'). Thus, this is a significant, non-random temporal association. The three main 'temporal clusters' denoted as '3', '4', and '5' are described in the text.



**Figure 5**: Frequent patterns of changes in albuminuria states for male patients (denoted by '1') and female patients (denoted by '2'): (a) worsening over a year, from a 'Normo-Low' level of albuminuria-state to a 'Micro' value; the pattern is valid for 9% of the male patients and 15% of the female patients); (b) worsening from 'Normo-High' to 'Micro'; the pattern is valid for 22% of the male patients and 22% of the female patients; (c) a stable Micro-Micro pattern; (d) a worsening of the renal state, from 'Micro' to 'Macro' albuminuria; the pattern is valid for 10% of the male patients and 9% of the female patients; (e) a 'Macro'-'Macro' stable pattern; (f and g) combinations of micro-albuminuria and various values of HbA1C-state. The legend panel (denoted by 3) represents the color scheme of the pattern's components. HbA1C, hemoglobin A1C.



albuminuria, HbA1C, and gender, in the first 1, 2, 3, or 4 years of follow-up.

The results (see table 3) show a clear trend of increasing predictive capabilities, as the duration of the period for collecting input data before the fifth year increases, up to using data from all of the previous 4 years, at which time the positive predictive value for micro/macro-albuminuria is 72.5%. Thus, the relative risk, given these features, increases by more than twofold versus the prior probability (30.2%) of this interesting class.

However, since the relevant features (patterns) can appear anywhere in the patient's temporal course, it is not necessarily clear (especially for clinical purposes) what is the relative risk for each patient after every year, starting from the beginning of the follow-up period for these patients.

Furthermore, it is also not clear from the TDM results what is the relative importance of each risk factor that was just discovered (eg, does each of the two quantitative risk factors have the same importance for men as for women?).

Thus, having identified at this point some of the risk factors, using the data-driven DM engine, in the next scenario we will examine the effect of integrating the albuminuria, HbA1C, and gender values. We will do that by switching back, within the ViTA-Lab framework, to the interactive, query-driven, visual mining capabilities to now ask even more specific queries regarding temporal associations across particular years of follow-up and specific risk factors.

#### Focusing on query-driven exploration

In the previous DM step, we found frequent patterns that are composed of various HbA1C-state and albuminuria-state values (see figure 5, panels f and g).

To continue our iterative analytical scenario, we would like to examine now what are the specific annual transition probability values, given different levels of HbA1C, creatinine, and albuminuria-state values (including the two normal subranges), into a value indicating a progressively deteriorating renal state (micro-albuminuria). We would also like to examine these transition probabilities for both male and female groups of patients. In particular, we would like to focus on the specific case of the association between the quantitative risk factors in the first year, and the target albuminuria-state values in the Table 3: The positive predictive value (PPV) of temporal patterns using the albuminuria-state values, HbA1C-state values, and gender, when the underlying raw data were measured continuously over two or more years, for the existence of either of the two classes of albuminuria-state values in year 5, and the corresponding area under the ROC curve (AUC)

Target classes (prevalence)	PPV, given the first year	PPV, given the first 2 years	PPV, given the first 3 years	PPV, given the first 4 years
'Normo-Low'/ 'Normo-High' Albuminuria-state values (69.8%)	70.3%	74.6%	78.7%	82.3%
'Micro'/'Macro' albuminuria-state values (30.2%)	53.8%	57.3%	64.1%	72.5%
AUC	0.558	0.677	0.753	0.815

HbA1C, hemoglobin A1C.

Table 4: Transition-probability values into a micro-albuminuria state in the fifth year with the effect of HbA1C

Year 1 albuminuria-state value	Year 1 HbA1C-state value	Probability of micro-albuminuria in year 5 for female patients (%)	Probability of micro-albuminuria in year 5 for male patients (%)
Normo-Low	Normal	6.33*	3.93
Normo-Low	High	11.63*	9.68
Normo-High	Normal	16.54	13.79*
Normo-High	High	15.73	21.62*

\*Denotes a significant difference between the proportion of patients who have the 'Normal' HbA1C-state value, and the proportion of patients who have the 'High' HbA1C-state value, for the same albuminuria-state value.

HbA1C, hemoglobin A1C.

fifth year of follow-up, for each gender. That is, can we predict micro-albuminuria 4 years ahead of time? Do the predictive factors vary between men and women?

Three different TACs were generated (separately for male and female patients, six charts in all) to examine the transition probabilities: albuminuria-state and HbA1C-state values during the first year to albuminuria-state 'Micro' value during the fifth year; similarly we examined the combination of albuminuriastate and creatinine-state values during the first year to albuminuria-state 'Micro' value during the fifth year; and finally, the combination of values of albuminuria-state, HbA1C-state, and creatinine-state during the first year to the 'Micro' albuminuria-state value during the fifth year. The selection criteria included having 5 years of data, and having as the earliest level of albuminuria-state either 'Normo-Low' or 'Normo-High'.

Due to the lack of space, the corresponding screenshots regarding HbA1C are represented in appendix A. We shall now summarize the results of the analysis.

# The association between HbA1C-state in the first year and micro-albuminuria in the fifth year

Table 4 summarizes the transition-probability values into a micro-albuminuria state in the fifth year of follow-up, resulting from our interactive visual exploration of the temporal pathways starting with either the 'Normo-Low' or 'Normo-High' albuminuria-state values in the first year.

A proportion test demonstrated that given a 'Normo-High' value of the albuminuria-state, there is a significant difference between men who had a 'Normal' HbA1C-state value compared with men who had a 'High' value of the HbA1C-state: Z = 1.612, p<0.05, but this difference did not exist for male patients who had a 'Normo-Low' value of the albuminuria-state during the first year.

However, in the case of female patients, there is significant difference, given a 'Normo-Low' value of the albuminuria-state, between having a 'Normal' HbA1C-state value and a 'High' value of the HbA1C-state: Z = 1.672, p<0.05. However, this

difference did not exist for female patients who had a 'Normo-High' value of the albuminuria-state during the first year.

It is interesting to note a proportion test does not show any significant difference in the portion of men versus women who display almost all of the four combinations of values of the albuminuria-state and creatinine-state, except for having the combination of a 'Normo-Low' value of the albuminuria-state and a 'Normal' HbA1C-state value: Z = 2.446, p < 0.05.

To sum up. A brief exploration, exploiting the temporal patterns provided by the data-driven TDM engine, augmented by some additional user-driven visual analysis, makes it clear that 'High' HbA1C levels might be a significant risk factor, in the case of male patients, for progressing into a micro-albuminuria state in the fifth year. This risk exists for patients who had in the first year albuminuria values that are usually considered by clinicians as being within the normal range.

# The association between first year creatinine-state and fifth year micro-albuminuria

Table 5 summarizes the transition-probability values into a micro-albuminuria state in the fifth year of follow-up starting with the two 'Normal' albuminuria-state values in the first year.

Recall that the three normal-range values, 'Normo-Low', 'Normo-High', and 'Normo-Severe' are usually considered by physicians as 'Normal' values of creatinine.

For either male or female patients who had the 'Normo-High' value of the albuminuria-state during the first year, a proportion test did not show a difference in the proportion of transitions into a micro-albuminuria state in the fifth year between any pair of creatinine-state values, including the two extremes. However, given a 'Normo-Low' value of the albuminuriastate, there was a significant difference in both female and male patients between having a 'Normo-Severe' value compared with having a 'High' value of the creatinine-state: Z = 2.809, p < 0.005 and Z = 2.108, p < 0.05, for female and male patients, respectively. Furthermore, in the case of female patients there is also a significant difference between having a 'Normo-Low' and 'Normo-Severe' creatinine-state value: Z = 2.054, p < 0.05.

Interestingly, a proportion test did not show any significant difference in the portion of male versus females patients who display any of the eight combinations of values of albuminuria-state and creatinine-state.

## CONCLUSIONS

We have presented an advanced, iterative ViTA-Lab framework, integrating data-driven (the DM engine, augmented by the Pattern Explorer interface) and user-driven (a visual, interactive query-driven) analysis of time-oriented multivariate clinical data. Both methods capitalize on a KBTA pre-processing phase.

We have demonstrated the benefits of such a framework by visually exploring a database of 22 000 patients with diabetes. We have shown how an iterative query-driven/data-driven visual analytical process can provide quick intuitions, and even result in actual answers to important clinical questions.

Examples of such results include the discovery of typical temporal pathways representing the progression of several renal damage markers in patients with diabetes over time; the assessments of the differences in the relative risk for future micro-albuminuria, given several potentially predictive factors,

Table 5: Transition probabilities into a micro-albuminuria state in the fifth year, given the creatinine state

Year 1 albuminuria-state value	Year 1 creatinine-state value	Probability of micro-albuminuria in year 5 for female patients (%)	Probability of micro-albuminuria in year 5 for male patients (%)
Normo-Low	Normo-Low	5.09*	7.98
Normo-Low	Normo-High	9.22	5.90
Normo-Low	Normo-Severe	9.84 <sup>*†</sup>	10.13†
Normo-Low	High	28.57†	17.65†
Normo-High	Normo-Low	14.66	14.93
Normo-High	Normo-High	13.82	18.96
Normo-High	Normo-Severe	20.27	24.00
Normo-High	High	32.14	29.73

\*Denotes the significant difference between the proportion of patients who have the 'Normo-Low' creatinine-state value, and the proportion of patients who have the 'Normo-Severe' creatinine-state value, for the same albuminuria-state value.

†Denotes the significant difference between the proportion of patients who have the 'Normo-Severe' creatinine-state value, and the proportion of patients who have the 'High' creatinine-state value, for the same albuminuria-state value.

among the two genders; the increasing, over 5 years of followup, positive predictive value of the temporal patterns discovered in a data-driven fashion within the patients' longitudinal data, whose components included the albuminuria-state, HbA1C-state, and creatinine-state values; and the observation of an increased relative risk for renal damage in the fifth year of follow-up, given the higher ranges of seemingly normal albuminuria and normal creatinine raw values, or higher than normal HbA1C-state values, starting even from the very first year of follow-up.

We have only briefly explored the diabetes nephropathy domain; additional potentially relevant concepts might complement our analysis, such as electrolytes, lipid values, etc.

Due to the generality of our approach, which only requires access to a relevant domain-specific temporal-abstraction ontology to interpret time-stamped raw data and exploit them for analysis purposes, we can also apply the same integratedanalysis approach to multiple additional clinical time-oriented domains. Indeed, we had previously created and used several domain-specific ontologies in multiple clinical domains, such as oncology, diabetes, monitoring children's growth, prenatal monitoring,<sup>10–15</sup> and even in non-medical domains, such as in the information security domain, for example when interpreting the time-oriented data of electronic devices to detect potential malware and to analyze the data of multiple devices.<sup>15,16</sup> These ontologies and the abstractions generated by applying them to data in these domains can support an integrated query-driven/data-driven temporal DM process similar to the one demonstrated in the current study.

Thus, we suggest that the ViTA-Lab framework can potentially serve as a virtual laboratory for investigations of large masses of longitudinal clinical databases, for the discovery of new knowledge through interactive exploration, clustering, classification, and prediction.

# ACKNOWLEDGEMENTS

We thank Robert Moskovitch for useful comments during the use and enhancement of the KarmaLego module, and for the inspiration provided by his classification-oriented discretization methods. We also would like to thank Drs Eitan Rubin and Michal Gordon for making the clinical data accessible for this research.

## **CONTRIBUTORS**

All three authors contributed significantly to the design of the methodology and the experiments and to the analysis of the data, as well as to the authorship of the manuscript itself. DK is the designated guarantor and corresponding author.

#### FUNDING

Part of this research was supported through the EU 7th Framework project, MobiGuide, EU award No. 287811.

## **COMPETING INTERESTS**

None.

## **ETHICS APPROVAL**

The Soroka Medical Center (the medical center associated with Ben Gurion University) IRB.

## **PROVENANCE AND PEER REVIEW**

Not commissioned; externally peer reviewed.

#### REFERENCES

- Remuzzi G, Schieppati A, Ruggenenti P. Nephropathy in patients with type 2 diabetes. *N Engl J Med.* 2002;346: 1145–1151.
- Hovind P, Tarnow L, Rossing P, *et al.* Predictors for the development of microalbuminuria and macroalbuminuria in patients with type 1 diabetes: inception cohort study. *BMJ*. 2004;328:1105.
- 3. Keim D, Mansmann F, Oelke D, *et al. Visual analytics: combining automated discovery with interactive visualizations. Discov Sci*, 2008:2–14.
- Keim D, Kohlhammer J, Ellis G, Mansmann F. eds. Mastering the Information Age Solving Problems with Visual Analytics. Goslar: Eurographics, 2010.
- Stacey M, McGregor C. Temporal abstraction in intelligent clinical data analysis: a survey. *Artif Intell Med.* 2007;39: 1–24.
- Shahar Y. A framework for knowledge-based temporal abstraction. *Artif Intell*. 1997;90:79–133.
- Boaz D, Shahar Y. A distributed temporal-abstraction mediation architecture for medical databases. *Artif Intell Med.* 2005;34:3–24.
- Lin J, Keogh E, Lonardi S, *et al.* A symbolic representation of time series, with implications for streaming algorithms. Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery; San Diego, California, 2003.
- Mörchen F, Ultsch A. Optimizing time series discretization for knowledge discovery. Proceedings of KDD05; Chicago, Illinois; 2005.
- Martins S, Shahar Y, Goren-Bar D, et al. Evaluation of an architecture for intelligent query and exploration of time-oriented clinical data. Artif Intell Med. 2008;43:17–34.
- Shahar Y, Das AK, Tu SW, *et al.* Knowledge-based temporal abstraction for diabetic monitoring. Proceedings of the Eighteenth Annual Symposium on Computer Applications in Medical Care; Washington, DC, 1994:697–701.
- Shahar Y, Musen MA. Knowledge-based temporal abstraction in clinical domains. *Artif Intell Med.* 1996;8:267–298.
- Kuilboer MM, Shahar Y, Wilson DM, *et al.* Knowledge reuse: temporal-abstraction mechanisms for the assessment of children's growth. Proceedings of the Seventeenth Annual Symposium on Computer Applications in Medical Care; Washington, DC, 1993:449–453.
- Shalom E, Friedman I, Shahar Y, et al. Towards a realistic clinical-guidelines application framework: desiderata, applications, and lessons learned. Proceedings of the 3rd International Workshop on Knowledge Representation for Health Care (KR4HC'11); Bled, Slovenia, 2011.

Published in: Lenz, R., Miksch, S., Peleg, M. *et al.* (eds). *Process Support and Knowledge Representation in Health Care.* Revised selected papers from the Proceedings of the BPM 2012 Joint Workshop, ProHealth 2012/KR4HC 2012, Tallinn, Estonia. Lecture Notes in Artificial Intelligence 7738, Springer, Berlin, 2013, pp. 56–70.

- 15. Shabtai A, Fledel Y, Elovici Y, *et al.* Using the KBTA method for inferring computer and network security alerts from time-stamped, raw system metrics. *J Comput Virol.* 2010; 6:239–259.
- Shabtai A, Shahar Y, Elovici Y. A distributed architecture for efficient parallelization and computation of knowledgebased temporal abstractions. *J Intell Info Syst.* 2012;39: 249–286.
- 17. Plaisant C, Mushlin R, Snyder A, *et al.* LifeLines: using visualization to enhance navigation and analysis of patient records. American Medical Informatic Association Annual Fall Symposium; 1998:76–80.
- Wang T, Plaisant C, Quinn A, *et al.* Aligning temporal data by sentinel events: discovering patterns in electronic health records. SIGCHI Conference on Human Factors in Computing Systems; 2008.
- 19. Falkman G. Information visualisation in clinical odontology: multidimensional analysis and interactive data exploration. *Artif Intell Med.* 2001;22:133–158.
- 20. Chittaro L, Combi C, Trapasso G. Visual data mining of clinical databases: an application to the hemodialytic treatment based on 3D interactive bar charts. Proceedings of Visual Data Mining VDM'2002; Helsinki, Finland, 2002.
- 21. Chittaro L. Information visualization and its application to medicine. *Artif Intell Med.* 2001;22:81–88.
- 22. Rind A, Wang T, Aigner W, *et al.* Interactive information visualization for exploring and querying electronic health records: a systematic review. *Found Trends Hum Comput* Interact. 2013;5:207–298.
- 23. Aigner W, Miksch S, Schumann H, et al. Visualization of *Time-oriented Data*. Springer, New York, 2011.
- Aigner W, Federico P, Gschwandtner T, *et al.* Challenges of time-oriented data in visual analytics for healthcare. In Caban J, Gotz D. eds. Proceedings of the IEEE VisWeek Workshop on Visual Analytics in Healthcare 2012, 2012.
- 25. Zhang Z, Ahmed F, Mittal A, *et al.* AnamneVis: a framework for the visualization of patient history and medical diagnostics chains. IEEE VisWeek Workshop; 2011.
- Zhang Z, Gotz D, Perer A. Interactive visual patient cohort analysis. In: Caban J, Gotz D. eds. Proceedings of the IEEE VisWeek Workshop on Visual Analytics in Healthcare— VAHC-2012, 2012.
- Brodbeck D, Degen M, Walter A. Masterplan: a different view on electronic health records. In: Caban J, Gotz D. eds. Proceedings of the IEEE VisWeek Workshop on Visual Analytics in Healthcare—VAHC-2012, 2012.
- Meyer T, Monroe M, Plaisant C, *et al.* Visualizing patterns of drug prescriptions with EventFlow: a pilot study of Asthma medications in the military health system. Proceedings of the Workshop on Visual Analytics in Healthcare—VAHC-

2012 in Conjunction with AMIA 2013; Washington, DC, 2013.

- 29. Wongsuphasawat K, Alexis Guerra Gómez J, Plaisant C, *et al.* LifeFlow: visualizing an overview of event sequences. Proceedings of the 2011 Annual CHI'11 Conference, 2011: 1747–1756.
- Wongsuphasawat K, Gotz D. Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization. *IEEE Trans Vis Comput Graph.* 2012;18: 2659–2668.
- 31. Gotz D, Wongsuphasawa K. Interactive intervention analysis. AMIA Annual Symposium; 2012:274–280.
- Perer A, Gotz D. Data-driven exploration of care plans for patients. CHI '13 Extended Abstracts on Human Factors in Computing Systems on—CHI EA '13, 439; 2013. doi:10. 1145/2468356.2468434.
- Perer A, Wang F. Frequence: interactive mining and visualization of temporal frequent event sequences. ACM Conference on Intelligent User Interface (IUI); 2014.
- 34. Allen JF. Maintaining knowledge about temporal intervals. *Commun ACM*. 1983;26:832–843.
- 35. Meyer T, Bennett P, Nelson R. Podocyte number predicts long-term urinary albumin excretion in Pima Indians with Type II diabetes and microalbuminuria. *Diabetologia*. 1999; 42:1341–1344.
- Aljumah A, Ahamad M, Siddiqui M. Application of data mining: diabetes health care in young and old patients. *J King Saud Univ—Comput Info Sci.* 2013;25:127–136.
- Vergouwe Y, Soedamah-Muthu S, Zgibor J, *et al.* Progression to microalbuminuria in type 1 diabetes: development and validation of a prediction rule. *Diabetologia*. 2010;53:254–262.
- Cho B, Yu H, Kim K, *et al.* Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods. *Artif Intell Med.* 2008; 42:37–53.
- 39. Huang Y, McCullagh P, Black N, *et al.* Feature selection and classification model construction on type 2 diabetic patients' data. *Artif Intell Med.* 2007;41:251–262.
- Hovind P, Tarnow L, Rossing P, *et al.* Predictors for the development of microalbuminuria and macroalbuminuria in patients with type 1 diabetes: inception cohort study. *BMJ* 2004;328:1105.
- 41. Hatsek A, Shahar Y, Taieb-Maimon M, *et al.* A scalable architecture for incremental specification and maintenance of procedural decision-support knowledge. *Open Med Inform J.* 2010;4:255–277.
- 42. Tao C, Wei WQ, Solbrig HR, *et al.* CNTRO: a semantic web ontology for temporal relation inferencing in clinical narratives. *AMIA Annu Symp Proc.* 2010;2010:787–791.
- Kaiser K, Miksch S. Versioning computer-interpretable guidelines: semi-automatic modeling of 'Living Guidelines' using an information extraction method. *Artif Intell Med.* 2009;46:55–66.
- 44. German-Shahar E, Leibowitz A, Shahar Y. An architecture for linking medical decision-support applications to clinical

databases and its evaluation. *J Biomed Inform.* 2009;42: 203–18. Reprinted in: Haux, R. (ed), *Yearbook of Medical Informatics 2009.* Stuttgart: F.K. Schattauer and The International Medical Informatics Association.

- Peleg M, Keren S, Denekamp Y. Mapping computerized clinical guidelines to electronic medical records: knowledge-data ontological mapper (KDOM). *J Biomed Inform*. 2008;41:180–201.
- Klimov D, Shahar Y, Taieb-Maimon M. Intelligent selection and retrieval of multiple time-oriented records. *J Intell Info Syst.* 2010;35:261–300.
- 47. Klimov D, Shahar Y, Taieb-Maimon M. Intelligent visualization and exploration of time-oriented data of multiple patients. *Artif Intell Med.* 2010;49:11–31.
- Klimov D, Shahar Y, Taieb-Maimon M. Intelligent interactive visual exploration of temporal associations among multiple time-oriented patient records. *Methods Inform Med.* 2009; 48:254–262.
- 49. Kaminski A, Klimov D, Shahar Y. Dynamic exploration and analysis of temporal interrelations between medical concepts for multiple time-oriented patient records. *Intelligent Data Analysis in Biomedicine and Pharmacology* 201. Bled, Slovenia, 2011.

- Moskovitch R, Shahar Y. Fast time intervals mining using the transitivity of temporal relations, knowledge and information systems. *Knowledge and Information Systems* First Published Online 27 December 2013. doi:10.1007/s10115-013-0707-x.
- 51. Moskovitch R, Shahar Y. Classification of multivariate time series via temporal abstraction and time intervals mining. *Knowledge and Information Systems* First Published Online 1 October 2014.
- 52. Shneiderman B. The eyes have it: a task by data type taxonomy for information visualizations. VL 1996:336–343.
- Moskovitch R. Classification of multivariate time series via learning of temporal abstraction patterns. PhD Dissertation, Department of Information Systems Engineering, Ben Gurion University. 2011.

# APPENDIX A THE ASSOCIATION BETWEEN HBA1C-STATE IN THE FIRST YEAR AND MICRO-ALBUMINURIA IN THE FIFTH YEAR

Exploration of several different temporal pathways. Shown are transition probabilities for male (left panels 1, 2) and



female (right panels 3, 4) patients, given the 'Normo-High' value of albuminuria-state, and either the 'High' (two top panels) or the 'Normal' (two bottom panels) values of the HbA1C-state, both characterizing the patients during the first year, into a micro-albuminuria value of the albuminuria-state concept, during the fifth year of follow-up. The tooltips denoted by '1' to '4' show the probability of ending in the 'Micro' value of the albuminuria-state concept in the fifth year for each of the four possible starting conditions. While the transition probability for male patients is affected by the HbA1C-state value (14.17% vs 31.71%), it does not seem to

## AUTHOR AFFILIATION

Department of Information Systems Engineering, Faculty of Engineering Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel be significantly affected by it in the case of female patients (16.54% vs 15.73%).

In order to create the displayed Temporal Association Charts (TACs), we first selected the 'Normo-High' value in the albuminuria-state concept (the first element from the left in TAC); thus, irrelevant patients are filtered out. In the second step, we selected the 'High' value of the HbA1C-state concept (the second element from the left in TAC) to get the top two panels for male and female patients respectively. The bottom two panels are obtained similarly by selection of the 'Normal' value of the HbA1C-state concept.