# Visual Data Mining of Clinical Databases: an Application to the Hemodialytic Treatment based on 3D Interactive Bar Charts

Luca Chittaro[1], Carlo Combi[2], Giampaolo Trapasso[1]

[1] HCI Lab, Dept of Math and Computer Science,
University of Udine,
via delle Scienze 206, 33100 Udine, Italy
chittaro@dimi.uniud.it
[2] Department of Computer Science,
University of Verona,
strada le Grazie 15, 37134 Verona, Italy
combi@sci.univr.it

**Abstract.** The capability of interactively mining clinical databases is an increasingly urgent need. This paper considers a relevant medical application (i.e., hemodialysis) and proposes a system for the visualization and visual data mining (VDM) of the collections of time-series acquired during hemodialytic treatments. Our proposal adopts bar charts as the basic visualization technique (because it is very familiar for clinicians) and augments them with several interactive features, exploiting a 3D space to significantly increase both the number of time-series that can be simultaneously analyzed in a convenient way and the number of values associated with each series.

## 1 Introduction

The capability of interactively mining patient clinical information is an increasingly urgent need in the clinical domain, due to the continuous growth in the number of parameters that can be automatically acquired and in the size of the databases where they accumulate [5]. This is particularly critical for the success of medical research projects which generate massive databases of patient data.

Some techniques for visual data mining (VDM) of multidimensional clinical databases are illustrated in [7]. They are mainly based on 3D versions of *parallel coordinate plots*. Graphical connections between points in adjacent planes are drawn in such a way that each patient's case is visually represented by a line connecting individual points referring to it. This allows for VDM of interesting patterns (e.g., a group of patients with the same profile results in parallel lines).

A different approach is presented by [10] and is based on *tables* displaying records of the clinical database and their attributes in highly compressed format such that they fit onto the screen. Users directly manipulate the table (e.g., performing zoom and filter operations) that dynamically rearranges itself. To compress the tables, the system relies on visualization criteria such as (i) neighboring cells with identical values are combined into a larger cell, or (ii) if there is no space to display a numeric value in its cell, the value is substituted by a small horizontal line whose position indicates relative size.

This paper explores a third possibility, especially suited to clinical databases containing time-series data. Since, historically, *bar charts* are a widely adopted approach to display time-series and are a very familiar representation for clinicians, we chose them as the basis of our visual approach. Unfortunately, while a bar chart allows for an easy comparison among the data values for a single time-series, when the considered task requires to compare a *collection* of time-series (such as a monitored signal from the same patient in different sessions of the same clinical test or treatment), traditional bar charts (as other historical approaches) become unfeasible. Therefore, we augment bar charts exploiting a 3D space and adding several interactive features. A 3D space can significantly increase both the number of time-series that can be simultaneously analyzed in a convenient way and the number of values associated with each time-series, but poses well-known problems such as occlusions, 3D navigation, difficulties in comparing heights, proper use of space, and the need for effective interaction techniques to aid the user in the analysis of large datasets (e.g., highlighting interesting patterns, checking trends,…). The limited capabilities of commercial tools that generate 3D bar charts have led well-known researchers (e.g. [9]) to classify these visualizations as "chartjunk 3D". However, solutions to the problems of 3D bar charts are emerging from research: e.g., Cichlid offers temporal animation capabilities of 3D stacked bar charts [2], while ADVIZOR allows one to interactively link the 3D bar chart representation with related 2D representations, compare heights with a "water level" plane (perpendicular to the bars) and use filtering tools [6].

Alternative approaches to time-series visualization have been recently proposed, e.g. drawing the timeline along spiral structures [12] is reported to allow for an easier detection of cyclic phenomena. However, we preferred to adopt bar charts, because they were familiar to clinicians. Moreover, we did not have a focus on a specific pattern such as cycles.

In the following, we first introduce the real-world clinical context we are working in and motivate the need for VDM in that context. Then, we illustrate the system we have built and its main features. Finally, we show some examples of how our system is being applied to the clinical context.

## 2 Hemodialysis and Visual Data Mining

Hemodialysis is the widely used treatment for patients with acute or chronic end-stage renal failure. During an hemodialysis session, the blood passes through an extra-corporeal circuit where metabolites (e.g., urea) are eliminated, the acid-base equilibrium is re-established, and water in excess is removed. In general, hemodialysis patients are treated 3 times a week and each session lasts about 4 hours.

Hemodialysis treatment is very costly and extremely demanding both from an organizational viewpoint [8] and from the point of view of the patient's quality-of-life. A medium-size hemodialysis center can manage up to 60 patients per day, i.e. more than 19000 hemodialytic sessions per year. Unfortunately, the number of patients that need hemodialysis is constantly increasing [12]. In this context, it is very important to be able to evaluate the quality of (i) each single hemodialysis session, (ii) all the sessions concerning the same patient, and (iii) sets of sessions concerning a specific hemodialyzer device or a specific day, for the early detection of problems in the quality of the hemodialytic treatment.

Modern hemodialyzers are able to acquire up to 50 different parameters from the patient (e.g., heart rate, blood pressure, weight loss due to lost liquids,…) and from the process (e.g., pressures in the extra-corporeal circuit, incoming blood flow,…), with a configurable sampling time whose lower bound is 1 sec. As an average example, considering only 25 parameters with a sampling time of 30 seconds, 12000 values (4*120*25) are collected in each session, and a medium-sized center collects more than 228 millions of values per year (considering 19000 provided treatments).

While the daily accumulation of huge amounts of data prompts the need for suitable techniques to detect and understand relevant patterns, hemodialysis software is more concerned with acquiring and storing data, rather than visualizing and analyzing it. Data mining applications can thus play a crucial role in this context. More specifically, *visual* data mining applications are of particular interest for three main reasons.

First, clinicians' abilities in recognizing interesting patterns are used suboptimally or not used at all in the current context. Visual mining of hemodialytic data would allow clinicians to take decisions affecting different important aspects such as therapy (personalizing the individual treatment of specific patients), management (assessing and improving the quality of care delivered by the whole hemodialysis centre), medical research (discovering relations and testing hypothesis in nephrology research).

Second, since data mining on the considered database is (at least, at initial stages) intrinsically vague for clinicians, the adoption of VDM techniques can be more promising than fully automatic techniques, because it supports clinicians in discovering structures and finding patterns by freely exploring the datasets as they see fit.

Third, the clinical context is characterized by a need for user interfaces that require minimal technical sophistication and expertise to the users, while supporting a wide variety of information intensive tasks. A proper exploitation of visual aspects and interactive techniques can greatly increase the ease of use of the provided solutions.

In summary, a clinical VDM system has to achieve two possibly conflicting goals: (i) offering powerful data analysis capabilities, while (ii) minimizing the number of concepts and functions to be learned by clinicians. In the following, we illustrate how our system attempts to achieve these two goals.

## 3 The Proposed Approach

The system we have built, called IPBC (*Interactive Parallel Bar Charts*) connects to the hemodialysis clinical database, produces a visualization that replaces tens of separate screens used in traditional hemodialysis systems, and extends them with a set of interactive tools that will be described in detail in this section.

Each hemodialysis session returns a time-series for each recorded clinical parameter. In IPBC, we visually represent each time-series in a bar chart format where the X axis is associated with time and the Y axis with the value (height of a bar) of the series at that time. Then, we layout the obtained bar charts side by side, using an additional axis to identify the single time-series, and we draw them in a 3D space, using an orthogonal view. It must be noted that also the additional axis has typically a temporal dimension, e.g. it is important to order the series by date of the hemodialysis session to analyze the evolution of a patient. An example is shown in Fig. 1, that illustrates a visualization of 50 time-series of 50 values each, resulting in a total of 2500 values (the axis on the right is the time axis for single sessions, while the axis on the left identifies the different time-series, ordered by date). Hereinafter, we refer to this representation as a *parallel bar chart*.

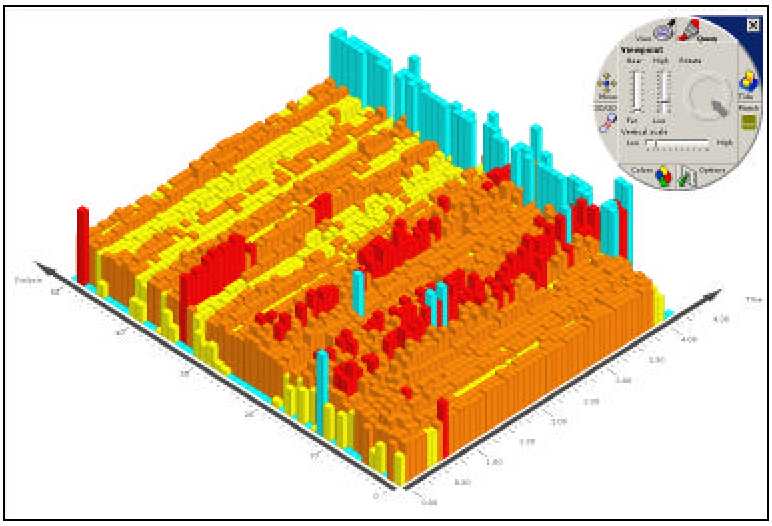

**Fig. 1.** A Parallel Bar Chart.

### 3.1. The RoundToolbar widget

In designing how the different interactive functions of IPBC should be invoked by the user, we wanted to face two different problems:

- First, one well-known limitation of many 3D visualizations is the possible waste of screen space towards the corners of the screen;

- Second, the traditional menu bar approach would require long mouse movements from the visualization to the menu bar and vice versa.

To this purpose, we designed a specific round-shaped pop-up menu (see Fig. 2), called *RoundToolbar* (RT), that appears where the user clicks with the right mouse button. The RT can be easily positioned in the unused screen corners, thus allowing a better usage of the screen space (e.g., see Fig. 1) and a reduction of the distance between the visualization and the menu. Moreover, to further improve selection time of functions with respect to a traditional menu, the organization of modes in the toolbar is inspired by Pie Menus [3]: in particular, the main modes are on the perimeter of the RT, and when a mode is selected, the center of the RT contains the corresponding tools (which are immediately reachable by the user, who can also quickly switch back from the tools to a different mode).

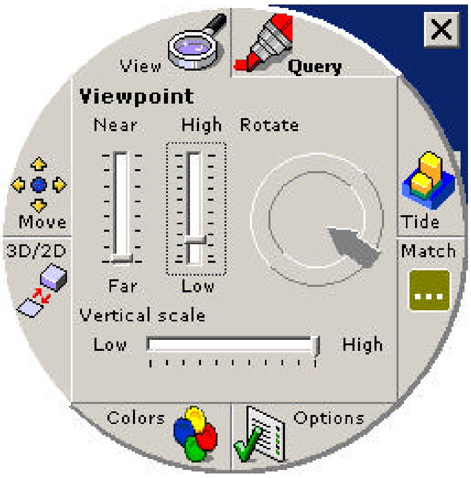


**Fig. 2.** Viewpoint mode.

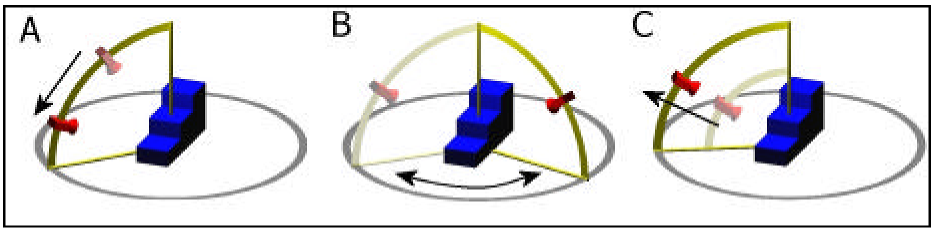


**Fig. 3.** Viewpoint movements: A) Low; B) Rotate; C) Far.

### 3.2. Changing Viewpoint

It is well-known that free navigation in a 3D space is difficult for the average user, because (s)he has to control 6 different degrees of freedom and can follow any possible trajectory. To make 3D navigation easier, when the *Viewpoint* mode is selected in the RT (as in Fig. 2), the proposed controls for viewpoint movement (*Rotate*, *High-Low* and *Near-Far*) cause movement along limited pre-defined trajectories which can be useful to examine the visualization: in particular, Fig. 3 shows how viewpoint movement is constrained. The remaining *Vertical scale* control in the *Viewpoint* mode is used to scale the bars on the Y axis. Vertical scaling has been included in the *Viewpoint* mode, because it has been observed that when users scaled the bars, they typically changed the viewpoint as the following operation.

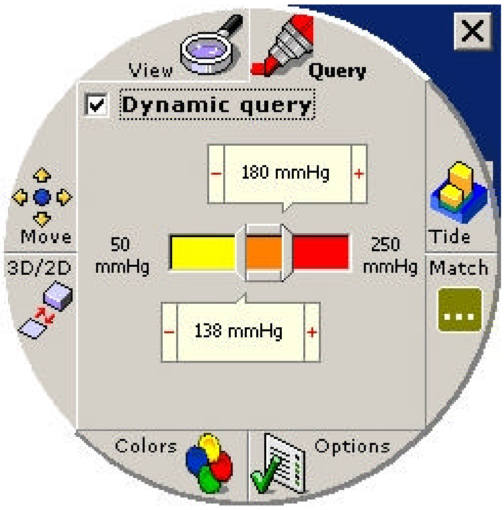


**Fig. 4. Dynamic Query mode.**

### 3.3. Dynamic Queries

IPBC uses color to classify time-series values into different ranges. In particular, at the beginning of a session, the user can define units of measure and her general *range of interest* for the values, specifying its lowest and highest value. These will be taken as the lower and upper bounds for an IPBC dynamic query control in the RT (as shown in Fig. 4) that allows the user to interactively partition the specified range into subranges of interest. Different colors are associated to the subranges and when the user moves the slider elements, colors of the affected bars in the IPBC change in real-time. Possible bars with values outside the specified general range of interest are highlighted with a proper single color. For example, Fig. 1 shows a partition that includes the three subranges corresponding to the colors shown by the slider in Fig. 4, and also some bars which are outside the user's predefined range. The color coding scheme can be personalized by the user with the *Colors* mode in the RT. The dynamic query control allows the user to:

- move the two slider elements *independently* (to change the relative size of adjacent subranges). For example, in Fig. 4, one has been set to 130 mmHg and the other to 180 mmHg. This can be done both by dragging the edges or (more easily) the tooltips which indicate the precise value. Plus and minus signs in the tooltips also allow for a fine tuning of the value.

- Move the two slider elements *together* by clicking and dragging the area between the two bounds. This can be particularly useful (especially when the other areas are associated to the same color), because it will result in a "spotlight" effect on the visualization: as we move the area, our attention is immediately focused on its corresponding set of bars, highlighted in the visualization.

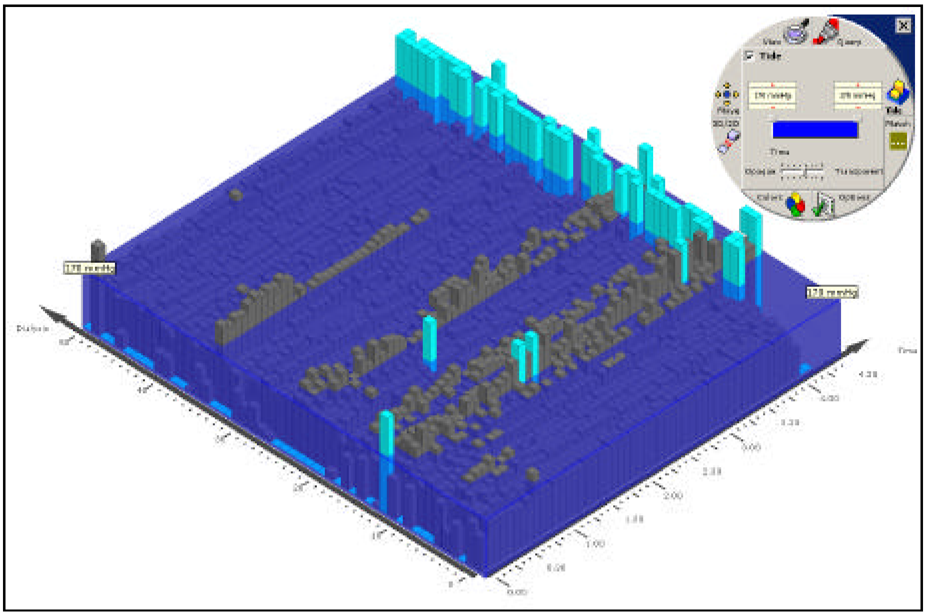

**Fig. 5.** Tide mode.

### 3.4. Comparing data with (time-varying) thresholds

A frequent need in VDM is to quickly perceive how many and which values are below or above a given threshold. This can be easily done with the previously described dynamic queries when the threshold is constant. However, the required threshold is often time-varying, e.g. one can be interested in knowing how many and which values are not consistent with an increasing or decreasing trend. For this need, IPBC offers a mode based on a tide metaphor. As it can be seen in Fig. 5, the *Tide* mode adds a semi-transparent solid to the visualization: the solid metaphorically represents a mass of water that floods the bar chart, highlighting those bars which are above the level of water. The slope of the top side of the solid can be set by moving two tooltips shown in the RT (that specify the initial and final values for the solid height), thus determining the desired linearly increasing or decreasing trend. The height of the solid can be also changed without affecting the slope by clicking and dragging the blue area in the RT. An *opaque/transparent* control allows the user to choose how much the solid should hide what is below the threshold. When the *Tide* mode is activated, all the bars in the user's range of interest are turned to a single color to allow the user to more easily perceive which bars are above or below the threshold; if multiple colors were maintained, the task would be more difficult, also because the chromatic interaction between the semitransparent surface and the parts of bars inside it adds new colors to the visualization.

The *Tide* mode can be also used to help compare sizes of bars by selecting a zero slope and changing the height of the solid (in this special case, *Tide* becomes analogous to the "water level" function of other visualization systems). Fig. 5 illustrates this latter case, while Fig. 9 shows a positive slope case.

Implementing a non-linear *Tide* would be relatively straightforward (only linear trends are anyway used by clinicians in the considered hemodialysis domain).

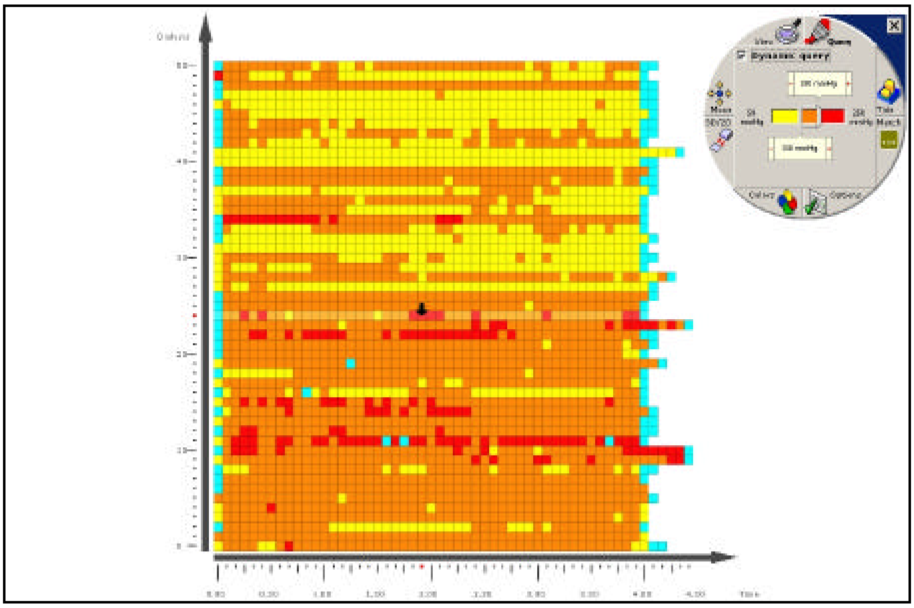

**Fig. 6.** Matrix Visualization.

### 3.5. Managing Occlusions

As any 3D visualization, IPBC can suffer from occlusion problems. To face them, the approach offers two possible solutions.

First, by clicking on the *2D/3D* label on the RT, the user can transform the parallel bar chart into a matrix format and vice versa. For example, Fig. 6 shows the same data as Fig. 1 in the matrix format. The transformation is simply obtained by automatically

moving the viewpoint over the 3D visualization (and taking it back to the previous position when the user deselects the matrix format). This can solve any occlusion problem (and the dynamic query control can still be used to affect the color of the matrix cells), but the information given by the height of the bars is lost. Transitions to matrix format and back are animated to avoid disorienting the user and allow her to keep her attention on the part of the visualization (s)he was focusing on.
Second, by directly clicking on any time-series in the 3D visualization, only the time-series which can possibly occlude the chosen one collapse into a flat representation analogous to the matrix one, as illustrated in Fig. 7.

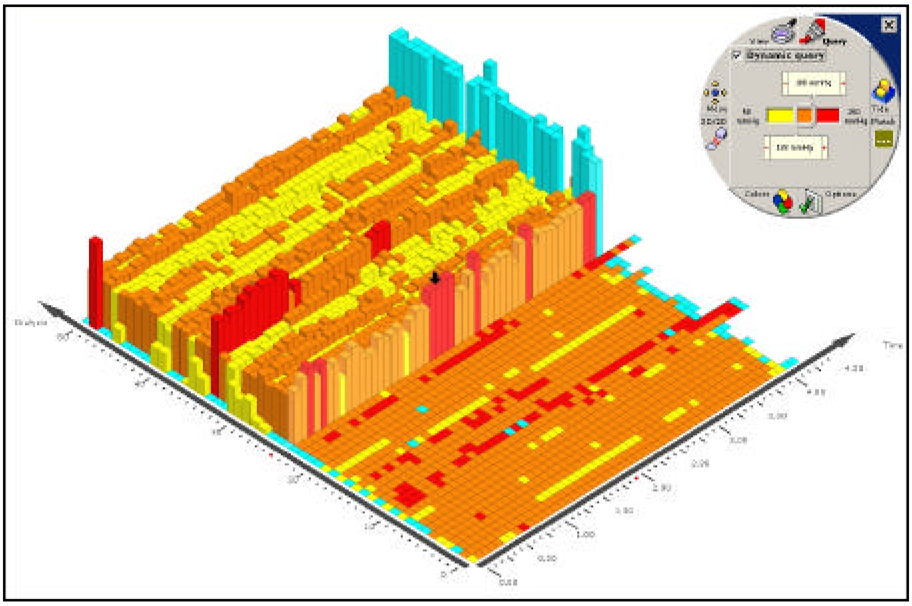


**Fig. 7.** Removing occlusions.

### 3.6. Pattern Matching

When the user notices an interesting sequence of values in one of the time-series, IPBC offers her the opportunity to automatically search for and highlight occurrences of a similar pattern in all the visualization (a detailed example will be described in Section 4.4).

The user selects her desired sequence of values in a time-series by simply dragging the mouse over it, then (s)he can specify how much precise the search should be, by indicating two tolerance values in the RT: (i) how much a single value can differ in percentage from the corresponding one in the given pattern, (ii) the maximum number (possibly zero) of values in a pattern that can violate the given percentage.

### 3.7. Mining Multidimensional Data

If multiple variables are associated to the considered time-series, IPBC can organize the screen into multiple windows, each one displaying a parallel bar chart for one of the variables. The visualizations in all the windows are linked together, e.g. if one selects a single time-series in one of the windows (or a specific value in a time-series), that time-series (or the corresponding value) is automatically highlighted in every other window. This (as some other features of IPBC) will be shown in more detail in the next section.

## 4. Mining Hemodialytic data

In the following, we will show how IPBC can be used during real clinical tasks, to help physicians evaluating the quality of the hemodialytic treatments given to single patients, on the basis of the clinical parameters acquired during the sessions. Each hemodialysis session returns a time-series for each parameter; different time-series are displayed side by side in the parallel bar chart according to date (in this case, the axis on the left chronologically orders the sessions).

The following examples are ordered according to the complexity of the related task: in particular, the first two tasks are relatively simple and are taken from the daily activity of clinicians, while the last two tasks are more complex and are performed by clinicians only in specific occasions (in the two considered examples, they are related to a detailed evaluation of the quality of care provided by nurses).

### 4.1. Mining patient signs data

A first task consists in analyzing patient signs, as the systolic and diastolic blood pressures and the heart rate; indeed, these parameters are important both for the health status of the patient and for the management of device settings during the hemodialytic session.

Let us consider, for example, the task of analyzing all the systolic pressures of a given patient: Fig. 8 shows a parallel bar chart (containing more than 5000 bars), representing the systolic pressure measurements (about 50 per session) during more than 100 hemodialytic sessions. In this figure, we can observe that the presence of out-of-scale values, usually related to measurement errors (e.g., the patient was moving; the measurement device was not properly operating), has been highlighted by specifying a proper range of interest (that highlights them in a suitable color) and hiding their height. In the specific situation represented in the figure, the presence of several out-of-scale values at the beginning of each session is due to the fact that nurses activate the measurement of patient's blood pressure with some delay with respect to the beginning of the session.

In the figure, the user is focusing on a specific session, avoiding occlusion problems (as described in Section 3.5). At the same time, with a dynamic query, (s)he is able to distinguish low, normal, and high blood pressures. In this case, the clinician can observe that the systolic pressure in the chosen session, after a period of low values (yellow bars), was in the range of normal values (orange bars). While the values for the chosen session correspond to a normal state, it is easy to observe that several sessions among those in the more recent half part of the collection contain several high values (in red) for the systolic blood pressure. Thus, the clinician can conclude that in those sessions the patient had some hypertension, i.e. a clinically undesired situation.

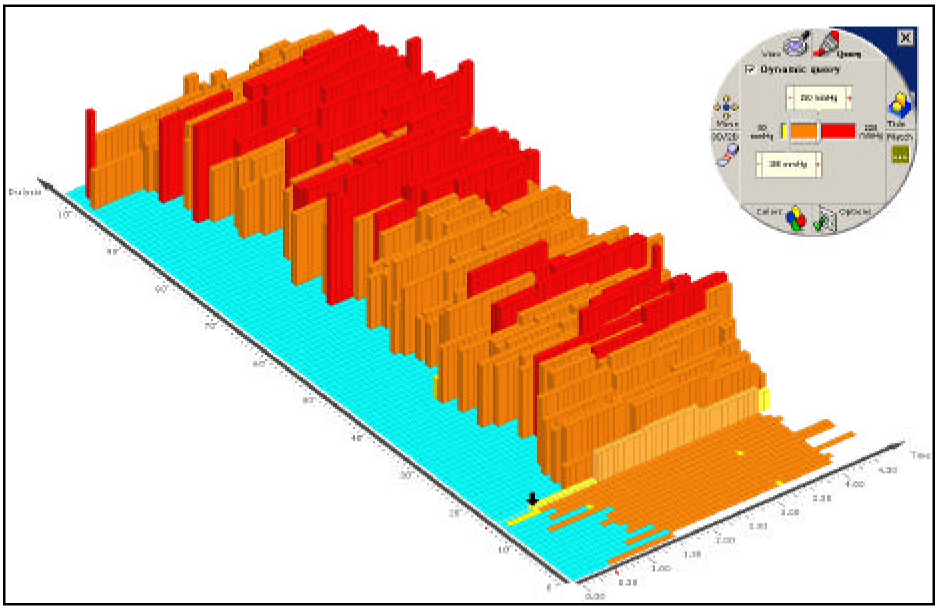

**Fig. 8.** Analyzing systolic blood pressures.

### 4.2. Mining blood volume data

Another task is related to observing the percentage of reduction of the blood volume during hemodialysis, mainly due to the removal of the water in excess. This reduction is sometimes slowed down to avoid situations in which the patient has too low blood pressures. In this case, VDM can benefit from the usage of the *Tide* mode. Fig. 9 shows an IPBC with more than 9000 bars, representing 36 hemodialytic sessions, containing about 250 values each. In this case, being the percentage of reduction of the blood volume increasing during a session, *Tide* allows the physician to distinguish those (parts of) sessions characterized by a percentage of reduction above or below the desired trend. In the figure, for example, the selected session has a first part emerging from the tide, while the last part is below. At the same time, it is possible to observe that one of the last sessions has the percentage of reduction above the tide during almost the entire session. The clinician can thus easily identify those (parts of) sessions with a satisfying reduction of the blood volume as the emerging (parts of) sessions.

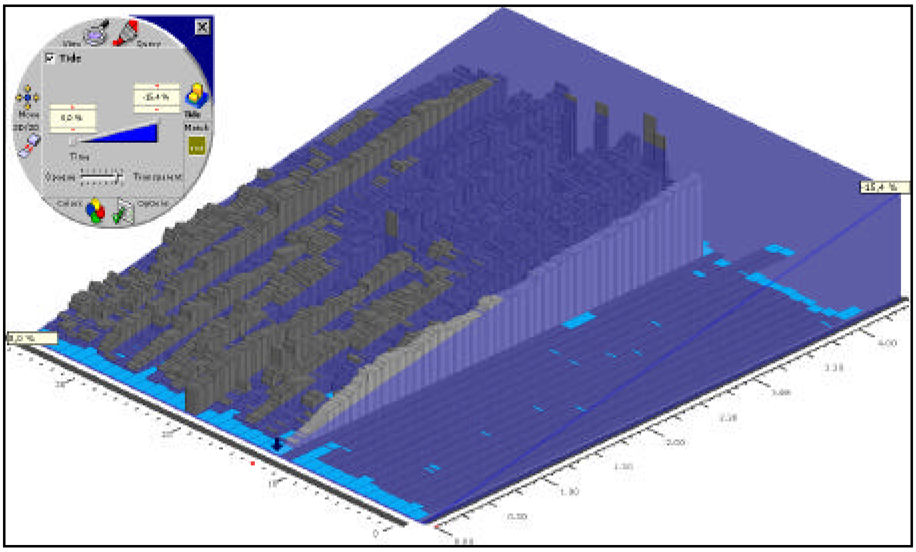

**Fig. 9.** Visualizing the time-varying reduction of the blood volume in the Tide mode.

### 4.3. Mining related clinical parameters

The next task we consider is related to the analysis of three related parameters: the systolic and diastolic blood pressures (measured on the patient) and the blood flow (QB) entering the hemodialyzer. QB is initially set by the hemodyalizer, but it can be manually set (reduced) by nurses when the patient's blood pressures are considered too low by the medical staff. It is thus interesting to visually relate QB and blood pressures, to check whether suboptimal QBs are related to low pressures. Otherwise, suboptimal values of QB would be due to human errors during the manual setting of the hemodialyzer. Fig. 10 shows the coordinated visualization of three clinical parameters for the same patient: the diastolic blood pressure (small window in the upper left part), the systolic blood pressure (small window in the lower left part), and QB (right window). The user can freely organize the visualization, switching the different charts from the smaller to the larger windows (by clicking on the arrow in the upper right part of the smaller windows). In the figure, the clinician is focusing on a session where the QB was below the prescribed value during the first two hours of hemodialysis (yellow color for QB) and (s)he has selected a specific value (the system highlights that value and the corresponding values in the other windows with black arrows). It is easy to notice that the suboptimal QB was related to low blood pressures (yellow bars in the corresponding time-series in the two small windows); then, QB was set to the correct value by nurses (see black arrow in the right window) only after blood pressures reached normal values (orange color in the corresponding charts). In this case, the physician can conclude that the suboptimal QB has been correctly set by nurses because of the patient's hypotension.

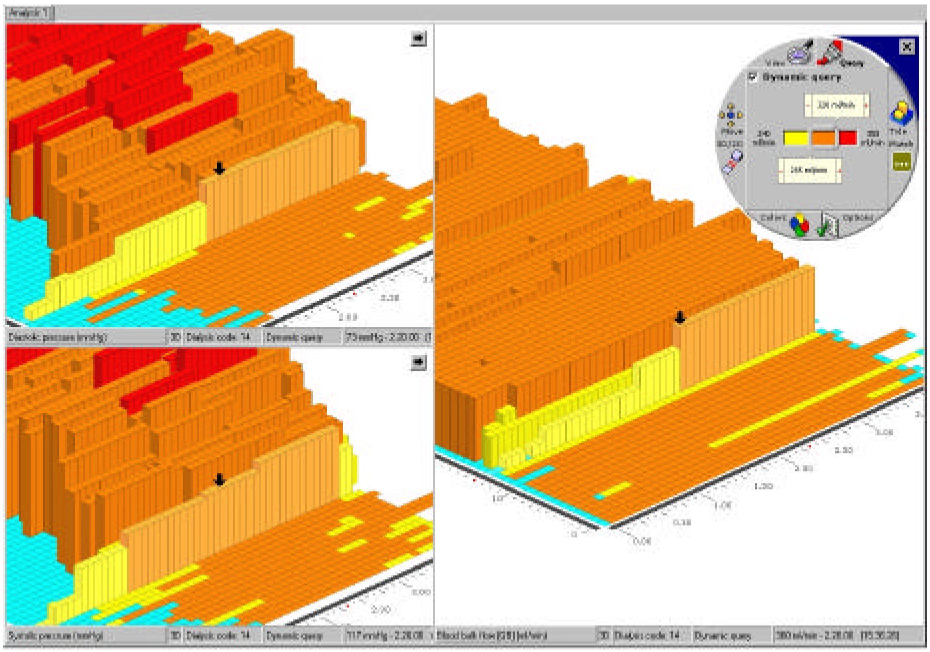

**Fig. 10.** Coordinated analysis of blood pressures and incoming blood flow.

#### 4.4. Mining for similar patterns

Finally, let us consider a task concerning the analysis of QB. As previously mentioned, the value of QB can be manually set by nurses and it may happen that this value is below the optimal one, due to hypotensive episodes. Fig. 11 shows a visualization where the clinician noticed a change of QB from a lower value to the correct one in a session: this means that, after a period of suboptimal treatment, the proper setting had been entered. Therefore, the clinician asks IPBC to identify QB patterns similar to the one (s)he noticed, by indicating it with the mouse, and setting the tolerance parameters (see Section 3.6). Fig. 11 shows the selected pattern (see the area near the black arrow) and the similar patterns automatically found by IPBC (two are in the lower right part of the figure, one in the upper left part): these patterns are identified by a line of a suitable color, which highlights the contours of the first and last bar of the pattern and intersects the inner bars. To avoid possible occlusion problems in visually detecting the patterns, the physician can move the viewpoint or switch to the matrix representation, where each pattern can be easily observed.

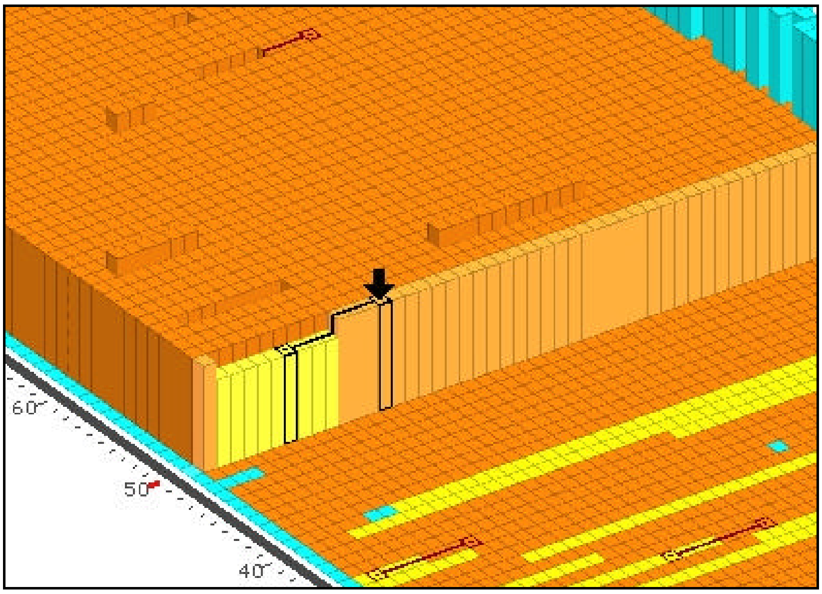


**Fig. 11.** Automatic Pattern Matching.

## 5. Conclusions and Future Work

In this paper, we described the main features of IPBC (Interactive Parallel Bar Charts), a VDM system devoted to interactively analyze collections of time-series, and showed its application to a real clinical database of hemodialytic data.

We are currently carrying out a field evaluation of IPBC with the clinical staff of the hemodialysis center at the Hospital of Mede, PV, Italy. One of the major advantages of IPBC that is emerging is that the visualization and its interactive features are very quickly learned and remembered by clinicians, the major disadvantage is that usage of screen space becomes difficult if a clinician tries to relate more than 3 collections of time-series simultaneously (Section 4.3 dealt with the analysis of 3 collections). This early feedback received from the field evaluation is helping us in identifying new research directions. Besides facing the problem of analyzing more than 3 collections in a convenient way, we aim to face another problem (that is considered very relevant by clinicians), i.e. dealing with time-series at different abstraction levels, allowing for both a fine exploration of time-series (e.g., to detect specific unusual values) and their coarse exploration (to focus on more abstract derived information). In both cases, we are working at the integration of parallel bar charts with other visualizations that can provide a synthetic view of data (e.g., the medical literature is proposing some computation methods to derive some quality indexes of the hemodialytic session from the time-series of that session). In particular, we are experimenting with Parallel Coordinate Plots, e.g. a trajectory in a plot could connect the quality indexes (typically, 5-7 values) of a session, and this high-level perspective would be linked to the much of a session, and this high-level perspective would be linked to the much more detailed perspective of the parallel bar chart.

**Acknowledgements**
This work is partially supported by a MURST COFIN 2000 project ("Analysis, Information Visualization, and Visual Query in Databases for Clinical Monitoring").

## References

1. Ahlberg, C., Williamson, C., Shneiderman B.: Dynamic queries for information exploration: An implementation and evaluation. Proc. of the CHI '92 Conference on Human Factors in Computing Systems, ACM Press, New York (1992) 619-626
2. Brown, J.A., McGregor, A.J., Braun H-W.: Network Performance Visualization: Insight Through Animation. Proc. of PAM2000: Passive and Active Measurement Workshop, Hamilton, New Zealand (2000) 33-41
3. Callahan, J., Hopkins, D., Weiser, M., Shneiderman, B.: An empirical comparison of pie vs. linear menus. Proc. of the CHI '88 Conference on Human Factors in Computing Systems, ACM Press, New York (1998) 95-100
4. Chittaro L. (ed.), Special issue on Information Visualization in Medicine, Artificial Intelligence in Medicine Journal, 22(2) (2001)
5. Chittaro L.: Information Visualization and its Application to Medicine, in [4] 81-88
6. Eick, S. G.: Visualizing Multi-Dimensional Data. ACM SIGGRAPH Computer Graphics, 34(1) (2000) 61-67
7. Falkman, G.: Information Visualization in Clinical Odontology: Multidimensional Analysis and Interactive Data Exploration, in [4] 133-158
8. McFarlane, P.A., Mendelssohn, D.C.: A call to arms: economic barriers to optimal hemodialysis care. Perit Dial Int 20 (2000) 7-12.
9. Shneiderman, B.: 3D or Not 3D: When and Why Does it Work?, invited talk at Web3D: 7th International Conference on 3D Web Technology, Tempe, AZ (2002)
10. Spenke, M.: Visualization and Interactive Analysis of Blood Parameters with Info-Zoom, in [4] 159-172
11. Tufte, E.R.: The Visual Display of Quantitative Information, Graphics Press (1982)
12. USRDS, The United States Renal data system, http://www.usrds.org
13. Weber, M., Alexa, M., Mueller, W.: Visualizing Time-Series on Spirals. Proc. of the IEEE InfoVis Symposium, IEEE Press, Los Alamitos, CA (2001)