Methods of Temporal Data Validation and Abstraction in High-Frequency Domains

Silvia Miksch[1], Andreas Seyfang[1],
Werner Horn[2, 3], Christian Popow[4], Franz Paky[5]

[1] Institute of Software Technology, Vienna University of Technology
{silvia, seyfang}@ifs.tuwien.ac.at
www.ifs.tuwien.ac.at/~{silvia, seyfang}
[2] Department of Medical Cybernetics and Artificial Intelligence, University of Vienna, werner@ai.univie.ac.at
[3] Austrian Research Institute for Artificial Intelligence, Vienna
[4] NICU, Department of Pediatrics, University of Vienna
popow@akh-wien.ac.at
[5] Department of Pediatrics, Hospital of Mödling
franz.paky@magnet.at

In medical domains, like Intensive Care Units both large amounts of on-line data and expert knowledge are available, but their automatic combination is hindered by practical concerns, like poor signal quality. In this chapter we present a set of methods to bridge this gap between erroneous high-frequency raw data and high-level symbolic representations. Interpretation of high-frequency data requires both time-oriented data validation and analysis resulting in high-level qualitative descriptions. The validation process consists of time-independent, time-point-, time-interval-, and trend-based methods to detect errors in the observed raw data as well as methods for its repair. The aim is to arrive at the most reliable data possible to obtain. The analysis abstracts different types of qualitative information concerning trends, values, and quality of the data. We developed various robust algorithms for both periodical and non-periodical curves to arrive at qualitative descriptions over time and to cope with artifacts in the data, which cannot possibly be detected in the previous validation steps.

1 Introduction

Most Intensive Care Units (ICUs) are well equipped with modern devices for patient monitoring. On-line recording of patient data and storage in

computer-based patient records (CPR) and patient data management systems (PDMS) become common-place in today's ICUs. Currently, the medical staff is suffering from information overloading caused by too many channels of on-line recording and from a vast amount of false alarms due to simple alarming policies (Lawless, 1994).

On the one hand, during the last years, several sophisticated knowledge-based monitoring and therapy-planning systems have been introduced (Uckun, 1994). These systems concentrate on optimizing data analyses and interpretation, on applying different kinds of accessible knowledge and information to enrich the reasoning process, and on minimizing manual data input by improving the technical equipment at modern clinics and by accessing computer-based patient records. However, particular time-oriented data-analysis methods are needed to cope with data in high-frequency domains and to ensure proper operation in life-threatening situations.

On the other hand, the monitors available at the ICUs are equipped with alarming systems, which can only detect obvious errors. These alarming systems apply simple methods of range checking, which are obviously too simple to be useful in a complex medical setting. As a result, the medical staff has a burdensome time to distinguish dangerous situations from false alarms. So, the supporting monitoring and therapy-planning systems are ineffective without error-detection methods because of the quite poor quality of the data (Carlson et al., 1993).

Therefore, we are aiming to overcome the problem of information overload and to improve the quality of data by applying the following four strategies:
First, we propose data validation methods to arrive at reliable data. The importance of data validation has been neglected in the past—the data received from the monitors is more faulty than is often realized (Gardner et al., 1992). Intensive efforts to detect artifacts require the combination of all information available, cross-validating various data sources, inspecting and reasoning about data points over time, and looking at trends to get a complete and consistent picture of the situation of the patient of the past and at present. In section we describe the methods for data validation and repair implemented in VIE-VENT, an open-loop knowledge-based monitoring and therapy planning system for artificially ventilated newborn infants (Miksch et al., 1993; 1996; Horn et al., 1997), which has been tested and evaluated in real clinical scenarios.

Second, data analysis methods are needed, which can handle time-oriented states and events, shifting contexts, and different expectations concerning the development of parameters. The process of this analysis is called *temporal data abstraction*. We describe the methods implemented in VIE-VENT and upcoming improvements in section . An advantage of using

these qualitative descriptions is their unified usability and interchangeability in further reasoning processes, regardless of the origin of the described data.

Third, a lot of non-systematic errors, called noise, can be eliminated by the data validation methods. However, not all errors can be detected. Therefore, the temporal data abstraction methods should be made less sensitive to such errors and at the same time provide information about the estimated quality of the data. In section we describe the calculation of a reliability score as a byproduct of the validation as implemented in VIE-VENT and in section we describe our approach to utilize statistical measures for the reliability of the data in the abstraction process.

Fourth, periodic high-frequency data call for a method to reason over changes in the form of the oscillations—not only its frequency and amplitude—in an intuitive way. Descriptions in terms of frequency spectrums or function matrices as used by popular approaches in the field of signal processing are not compatible with the representation physician use when describing curves. In section 4.3 we describe our ongoing research on this topic.

In section 2 we describe the need for data validation and abstraction and describe the characteristics of the used data. We present our approaches to preprocessing data (time-oriented data validation) in section 3 and those to time-oriented data abstraction in section 4. The evaluation and benefits of our approach are discussed in section 5 and future work in section 6.

2 Motivation - The Characteristics of the Data

2.1 The Need for Time-Oriented Data Validation

In the following we will motivate the necessity of effective data validation illustrating our experiences with medical on-line data from ICUs.

We evaluated on-line data sets obtained from newborn infants with various respiratory diseases. The data were collected from the monitoring system of a neonatal Intensive Care Unit (NICU) once per second (16-28 hours of continuous data recording for each newborn infant). The data sets consist of measurements of continuously assessed quantitative data (e.g. transcutaneous partial pressure of oxygen ($P_{tc}O_2$), the pulse frequency (*PULS*) given from pulsoximetry), discontinuously assessed quantitative data (e.g. ventilator settings like *PIP*, *PEEP*, results of invasive blood-gas analyses like *pH*, P_aO_2 where *a* denotes a measurement from arterial

blood), and continuously assessed qualitative data (e.g. clinical parameters like spontaneous breathing effort, chest wall extension).

Visualization and analysis of these data sets enabled a closer insight into the validity and the quality of the observed data, as well as the importance of secure and trustworthy data for further reasoning:

1. Small movements of the infant resulted in an unexpectedly high volume of data oscillation. This is specifically a problem of pulsoximetry. For example, small movements of the neonate result in sequences of unusable oxygen saturation (S_aO_2) measurements.
2. The measurements were frequently invalid caused by external events, which have to be performed regularly (e.g. calibration of transcutaneous sensors every three to four hours, scheduled endotracheal suctioning).
3. Continuously and discontinuously assessed measurements, which should reflect the same clinical context, frequently deviated from each other as a result of the individual situation of the patient or of variations in the environmental conditions under which the sensors operate.
4. Additional invalid measurements were caused by on-line transmission problems or were unexplainable.
5. Some errors occur because different people input data from in different environments and in different experimental settings.

Noisy and erroneous data is a serious problem—the data analysis and data mining methods should be made less sensitive to such non-systematic errors. In the machine learning literature the problem of noisy data has been expensively studied. On the one hand, when generating the rules from training data, the noise should be eliminated to make the rules more general and accurate. On the other hand, in some systems just the opposite is true: adding the noise to training data resulted in smaller misclassification of unseen examples ((Quinlan 1986) cited in (Cios, Pedrycz, & Swiniarski, 1998)). However, in our domain noise is seen as distracting from the real information and thus both data validation and abstraction must provide various methods to minimize the influence of noise on the outcome.

2.1.1 Related Work

Classical artifact-recognition methods mostly come from the field of statistical signal processing techniques and neural networks. Statistical signal processing, like Kalman filtering, is computationally expensive (Sittig & Factor, 1990). It puts much power in processing signals at a very low level, which may be unnecessary, if we know from high-level reasoning

processes that the signal is useless. The same arguments hold for artificial neural networks (Sittig & Orr, 1992).

Error detection is inevitable in anesthesia monitoring (van der Aa, 1990) and post-operative care (Sukuvaara et al., 1992). The combination of range checks and validation and invalidation rules has been successfully applied by (Garfinkel et al., 1989) to eliminate false alarms and at least range checking facilities are standard for today's monitors in ICUs. However, commonly used systems produce numerous false alarms—or, if switched off—missing alarms (Lawless, 1994).

Most methods used today concentrate on numerical methods and do not take into account the clinical context. These methods are successful for particular problem characteristics—detecting values, which are not within certain ranges and trend values, which are physiologically implausible. But they cannot classify data as unreliable, because a large portion of reliability checking is dependent on the correct interpretation of the clinical context. Further, cross-checking of different parameters needs a very high, abstract level of reasoning. Such a reasoning gives insight into the reliability of measured data, both on a specific data point and on the trend over some selected time period.

Avoidance of wrong alarms, reliable monitoring, and effective therapy planning requires data validation procedures, which combine numerical methods with validation methods operating on derived qualitative time-oriented descriptions of state and grade values and various combinations thereof.

2.2 The Need for Deriving Temporal Abstractions

Beside the quality of data, monitoring and therapy planning in real-world environments involves numerous other data analysis problems:

1. Long-term monitoring requires the processing of a huge volume of data generated from several (monitoring) devices and individuals.
2. The available data occur at various observation frequencies (e.g. high or low frequency data), at various regularities (e.g. continuously or discontinuously sampled data), and are of various types (e.g. qualitative or quantitative data).
3. A time-oriented analysis process has to cope with a combination of all these data sources.
4. The underlying domain knowledge about the interactions of parameters is vague and incomplete.
5. The interpretation context is shifting depending on observed data.

6. The underlying expectations regarding the development of parameters are different according to the interpretation context and to the degrees of the parameters' abnormality.

2.2.1 Related Work

Traditional theories of data analysis (Avent & Charlton, 1990; Kay, 1993) mostly deal with well-defined problems. However, in many real-world cases the underlying structure-function models or the domain knowledge and models are poorly understood or not applicable because of their complexity and because knowledge is often incomplete or vague. Therefore, in the medical domain statistical analysis, control theory, or other techniques are often unusable, inappropriate or at least only partially applicable (Miksch et al., 1996; Horn et al., 1997).

To overcome the mentioned limitations, time-oriented analysis methods were proposed to derive qualitative values or patterns of the current and the past situation of a patient (e.g. transcutaneous partial pressure of carbon dioxide ($P_{tc}CO_2$) is *slightly below the target range*, or $P_{tc}CO_2$ is *increasing*). These data analysis methods are referred as *data-abstraction methods*, a term originally introduced by Clancey in his classical proposal on heuristic classification (Clancey, 1985). *Temporal* data abstraction represents an important subgroup where the processed data are temporal. Atemporal data abstraction is substantially simpler than temporal abstraction, because time adds a new dimension and temporal dependencies dramatically increase the complexity of a problem.

An advantage of using such qualitative descriptions is their unified usability in the system model, regardless of their origin. Several significant and encouraging approaches have been developed in the past years.

Haimowitz et al. (1995) have developed the concept of trend templates ($TrenD_x$) to represent all the information available during an observation process. A trend template defines disorders as typical patterns of relevant parameters. These patterns consist of a partially ordered set of temporal intervals with uncertain end-points. Trend templates are used to detect trends in time-stamped data. The RÉSUMÉ project (Shahar & Musen, 1996) performs temporal abstraction of time-stamped data without predefined trends. The system is based on a knowledge-based temporal-abstraction method, which is decomposed into five sub-tasks: temporal context restriction, vertical temporal inference, horizontal temporal inference, temporal interpolation, and temporal pattern matching. Larizza et al. (1997) have developed methods to detect predefined courses in a time series. Complex abstraction allows to detect specific temporal relationships between intervals. The overall aim was to summarize the patient's behavior over a predefined time interval. Belazzi et al. (1999)

utilize Bayesian techniques to extract overall trends from cyclic data in the field of diabetes. Keravnou (1997) focuses on the periodicity of events derived from the patient history.

All these approaches are dealing with low-frequency data. Therefore, the problems of oscillating data, frequently shifting contexts, and different context-specific expectations of the development of parameters are not covered.

In the field of NICUs, Hunter et al. (1999) developed a tool to detect significant events like probe changes in recorded data from monitors. The algorithm is based on joining temporal intervals until the error of the linear regression calculated from the raw data points within that window exceeds a particular threshold. Although our approach utilizing a *spread* (see section 4.1 and (Miksch et al., 1999)) differs significantly, its development was inspired by this.

3 Preprocessing: Time-oriented Data Validation

The parts of the data abstraction methods described in section 4 are interwoven with the data validation process. First, the data validation process uses the numerical values of the parameters to arrive at reliable values which are transformed into unified qualitative descriptions by the data abstraction process. Second, it applies these derived qualitative descriptions to detect faulty measurements. The major aim of the data validation process is to detect faulty measurements or artifacts and finally to arrive at reliable measurements. An artifact is a situation where a measured variable does not reflect the clinical context.

We perform a two-step data validation process based on different temporal ontologies: first, a context-sensitive examination of the plausibility of input data and second, applying repair and adjustment methods for correcting wrong or ambiguous data. The final result is a classification of the input data as “correct”, “wrong”, “unknown”, or “adjusted”. A measurement is classified as “adjusted” if a “wrong” or “unknown” value is corrected by a repair or adjustment method. If a faulty measurement is recognized and no repair or adjustment method can be applied, the measurement is classified as “wrong”. If no data for a measurement from the monitor is received and no value could be estimated, then the measurement is classified as “unknown”. Otherwise it is classified as “correct”. Not all methods mentioned below lead to a final classification. Some of them (like, the time-point-based functional dependencies) result in an intermediate and ambiguous classification of “some are wrong”. This information is

forwarded to and handled by the repair and adjustment module which provides strategies for repairing and adjusting not plausible or missing values based on the same temporal ontologies as the data validation module.

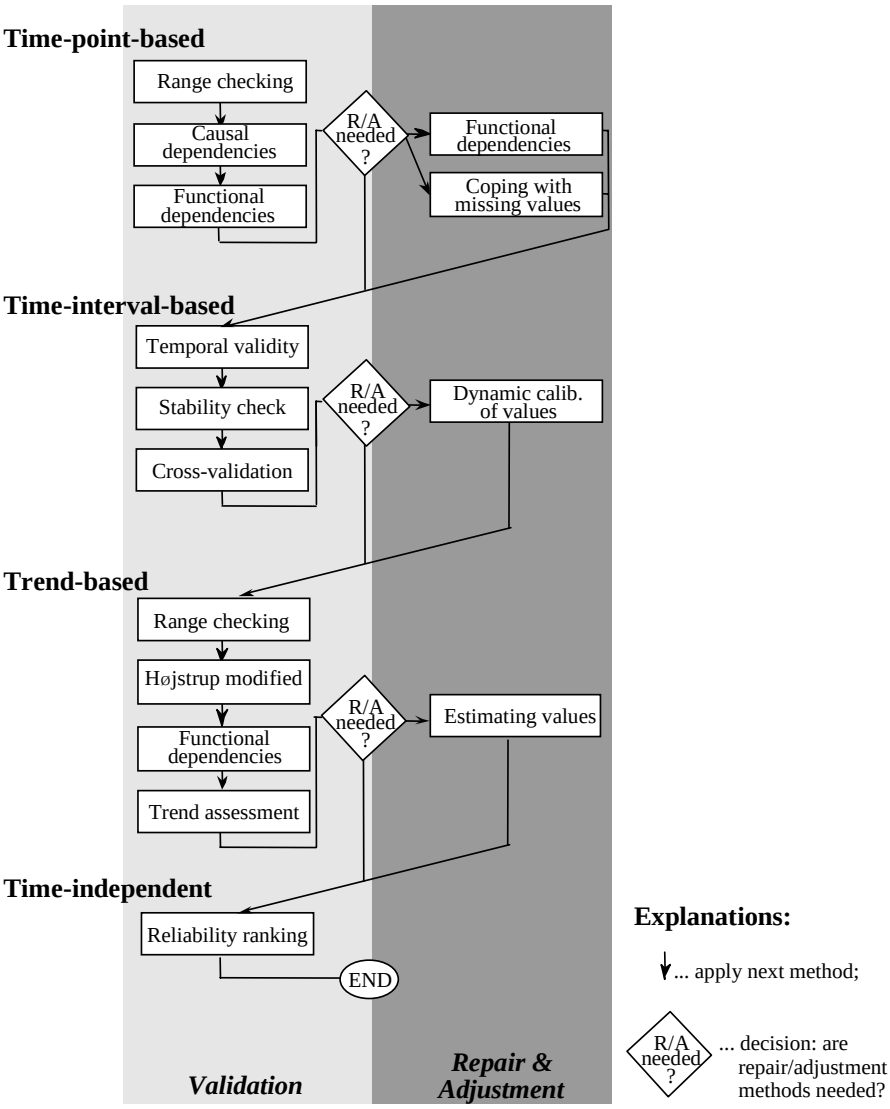

Figure: Overview and interaction of the components of the data validation and repair/adjustment modules. The left-hand side labels indicate the different temporal-based ontologies and the corresponding validation methods. On the right-hand side the possible repair or adjustment methods are mentioned.

We divide our methods into four types based on their underlying temporal ontologies: time-point-based, time-interval-based, trend-based, and time-independent validation and repair. Figure 1 gives an overview of the particular categories and their interactions. A detailed description of the whole process is given in (Horn et al., 1997).

3.1 Time-Point-Based Validation and Repair

The time-point-based category uses the value of a variable at a particular time point for the reasoning process. This concept can handle any kind of data. It benefits from the transparent and fast reasoning process but suffers from neglecting any information about the history of the observed parameters.

We distinguish the following validation and repair methods:

3.1.1 Validation: Range Checking

The range checking determines whether a quantitative value is within an acceptable range. It is simple but has shown very powerful to detect disconnections and missing measurements. Most modern ICU's equipment is able to perform range checks by itself.

We have enhanced this method by adding additional attributes, which define the clinical context (e.g. arterial, IPPV). There are look-up tables for all input parameters covering the plausible ranges. A parameter in the look-up table is specified by a parameter name, a list of attribute descriptors, an upper limit and a lower limit. For example, (CO_2, (arterial, IPPV), 10, 140), where "arterial" refers to the kind of blood gas analysis and IPPV to the mode of ventilation.

3.1.2 Validation: Causal Dependencies

Causal dependencies establish a relationship between different parameters. Qualitative values (e.g. *chest wall extension = small*) are related to numerical ranges of other parameters (e.g. *tidal volume < 5 ml/kg*). A causal dependency can be bidirectional—as shown in the example above—or unidirectional. In the bidirectional case we can only conclude that some of the parameters are wrong if the dependency is violated. The unidirectional case allows to invalidate a specific parameter. For example, S_aO_2 is invalidated if we cannot find a valid pulse (from pulsoximetry) or if we detect a substantial difference between the pulse and the heart rate from ECG (*HR*, measured in *beats*/*min*):

$$valid\ (PULS) = false \rightarrow valid\ (S_aO_2) = false \quad (1)$$

$$|\ HR - PULS\ | > 8 \rightarrow valid\ (S_aO_2) = false \quad (2)$$

Equation 2 can be used only if we have a valid *HR* and a valid *PULS*. In fact, such dependencies define an implicit ordering of parameters with respect to the application of validation procedures.

3.1.3 Validation and Repair: Functional Dependencies

Functional dependencies are useful for both numerical and qualitative parameters. Applying a functional dependency not only provides a mean for validating the parameters of the function, but also for repair of an invalid parameter.

Functional numerical dependencies are used to provide a value for a dependent parameter and to check inadequate data transmission for parameters where we know the exact functional relation.

Qualitative functional dependencies establish a relationship between derived qualitative values of different parameters. Due to the unified scheme for the qualitative values of all blood-gas measurements as shown in section it is easy to compare different measurements. For blood-gas measurements we expect that measures taken from different sites (arterial, venous, capillary, and transcutaneous) belong to the same qualitative data point region, or at least to the neighboring one. For example, we expect the same classification of the transcutaneous $P_{tc}CO_2$ and the invasive capillary P_cCO_2 measurements. If we detect, e.g. $P_{tc}CO_2$ is *substantially above the target range* and P_cCO_2 is *normal* we remember the ambiguity of the transcutaneous and the capillary carbon dioxide measurement. Which of the values is more plausible depends on the static reliability ranking discussed in section and the dynamic reliability score computed by each of the various validation methods. Later on in the validation process we will either invalidate one of the two measurements or repair it using dynamic calibration.

3.1.4 Repair: Coping with Missing Values

This method is applied if a value is marked as “unknown”, “wrong”, or “some are wrong” and if it could not be adjusted by any other method. There are two options to deal with missing values:

- *Simplified reasoning process.* This process uses only a few—most essential—parameters for further reasoning.

- *No solution.* When a critical situation has arisen in the past, no solution can be derived and the recommendations of appropriate treatments are delegated to the physicians.

Although not providing a solution might not seem to be a feature at first glance, for a system deployed in the medical domain it is vital to ensure that any output is based on solid grounds and that erroneous recommendations are strongly prohibited. The other advantage of a system “knowing its limits” is that operation is resumed immediately after input is valid again, without any need for an intervention by the user, as would be unavoidable for systems just “getting lost” in case of unexpected failure of input devices.

3.2 Time-Interval-Based Validation and Repair

The time-interval-based category of validation and repair deals with the values of different variables within a time interval.

We distinguish the following validation and repair methods:

3.2.1 Validation: Temporal validity

Temporal validity sets the time interval during which a parameter is valid. We distinguish two kinds of temporal validity according to the regularities of the sampled data.

1. For *discontinuously* assessed data there are two possibilities for setting the valid time interval:

   (a) The user can specify the duration of validity for each entered datum. E.g., "P_aO_2 should be valid for the next 30 minutes”.

   (b) For each parameter there is a predefined default maximum duration of validity.

   A discontinuously assessed parameter value looses its validity, if one of the following conditions becomes true:

   (a) the time interval of the parameter’s validity has elapsed,

   (b) a new value of the parameter is available, or

   (c) an external event (e.g. calibration of sensors) enforces to set the parameter invalid.

   The reliability score of a discontinuous parameter becomes smaller over time. For each parameter, a temporal validity interval is defined,

which determines how long the time-interval-based repair method *dynamic calibration* (see below) can be active.

2. *Continuously* assessed data are handled in a different way: instead of valid time intervals we define *invalid* time intervals. The user can set a parameter invalid explicitly, if specific external events take place (e.g. calibration of sensors, new application of sensors, disconnection).

3.2.2 Validation: Stability Check

After a period of invalidity of a parameter it is essential to enforce some (short) period of stability before the parameter is set back valid. This is specifically true for rapidly changing parameters like S_aO_2. The stability check defines allowed changes in the values of parameters. It compares the new value of a parameter with previously assessed values within a predefined time interval. This method is applicable for continuously assessed data only. We distinguish two situations:

1. Allowed changes of parameter values *without* a therapeutic action: The first value of a parameter, which is classified valid by all other validation methods becomes a candidate for stability testing. During time interval *n* we require, e.g.

$$\forall_{i,\, i=1,\ldots,n} : |\, S_aO_2(t) - S_aO_2(t+i)\,| \leq \varepsilon \qquad (3)$$

1. For excellent stability of S_aO_2 we currently use $n = 120sec$ and $\varepsilon = 5\%$. The effect of the stability check is a delay in setting a parameter valid again. E.g., for S_aO_2 we will wait 120 seconds until the data values can be used again. If the stability check succeeds, we are able to reuse the values of the last 120 seconds. This results in a recalculation of the trends.

2. Allowed changes of parameter values *after* a therapeutic action: we expect a particular parameter to improve towards the normal range after a certain delay time. Besides the fact that therapeutic actions are not recommended in case the guiding parameters are invalid, a stability check as defined above is less useful. A larger ε for the direction of the desired improvement is used in this case.

3.2.3 Validation: Cross-Validation

Cross-validation of data from different sources is the time-interval-based utilization of qualitative functional dependencies described in section 3.1. Its specific use is the correlation of a parameter *X* which gives a quite exact measurement but is rarely available with a parameter *Y* which is

inexact but available continuously. The basic assumption is that *X* behaves like *Y*.

As an example taken from ventilation management, *X* is an invasively measured blood gas and *Y* is a transcutaneous blood gas. If cross-validation detects a significant qualitative difference between, e.g. P_aCO_2 and $P_{tc}CO_2$ as described above, and both parameters are not invalidated by other methods, we apply dynamic calibration.

3.2.4 Repair: Dynamic Calibration

Dynamic calibration is a time-interval-based repair method, which repairs continuously assessed data values by applying a repair function which utilizes the difference between the discontinuously assessed data value *X* and the corresponding continuously assessed data value *Y*. This repair function adjusts the less reliable continuos value over a *temporal validity interval* to the reliable value of *X*. The resulting repaired value of *Y* receives a high reliability score which subsequently decreases over time. More details can be found in (Horn et al., 1997).

3.3 Trend-Based Validation and Repair

Trend-based validation analyzes the behavior of a variable during a time interval. A trend is a significant pattern in a sequence of time-ordered data. Therefore, the following methods can handle only continuously observed variables. They benefit from dynamically derived qualitative trend descriptions presented in section 4.4.

Based on physiological criteria, four kinds of trends of the time-stamped data samples can be discerned. They differ in the length of the sequence of data they use to calculate the trend. Further, they differ in the validity criteria for the determination of a valid trend. In monitoring more recent data are more important compared to older measurements. Thus we defined two criteria of validity to ensure that a trend is actually meaningful: (1) a certain minimum amount of valid measurements within the whole period, and (2) a certain amount of valid measurements during the last 20 percent of the time interval. These limits are defined by experts based on their clinical experience. They may easily be adapted to a specific clinical situation based on the frequency at which data arrives. Table 1 summarizes the trends and their criteria. For each kind of trend the actual growth rate and the derived qualitative trend category is determined as detailed in section 4.4.

kind of trend	sequence duration (minutes)	percentage of valid measurements for	
		whole sequence	last 20% of sequence
very short	1	50%	100%
short	10	40%	80%
medium	30	30%	60%
long	180	20%	40%

Table 1: Criteria of trend validity.

We distinguish the following validation and repair methods:

3.3.1 Validation: Range Check of the Growth Rate

A first basic check is the inspection of the growth rate. It is a sensible method for recognizing problems with the technical equipment, e.g. sensor loss. Range checks are applied on the very short-term trend and therefore react very fast.

3.3.2 Validation: Højstrup Method Modified

The modified Højstrup method recognizes growth rates, which are unacceptable after a certain amount of time. It recognizes implausible values by inspecting the temporal behavior of measurements. The temporal behavior is given as a function of measured values over time. Measurements are classified as implausible if the growth of this function is either too steep or the growth rate lies above a threshold and lasts for too long. The basic idea is given in (Højstrup, 1992). We have modified the algorithm to the needs of real-time monitoring in ICUs: the correlation function K is replaced by a measurement for the deviation of the last two points from the mean. We further may not assume a normal distribution of the differences. Therefore, the error threshold E is derived from knowledge about the maximum growth rate to accept and the desired rigidity of the system.

The algorithm is given in (Egghart, 1995; Horn et al., 1997). The main advantage of the method is the ability to select an area of growth between a value where it never signals an invalidity and a value where it immediately signals an invalidity. In between, the lower the growth rate the longer it will take to signal an invalidity.

3.3.3 Validation: Trend-Based Functional Dependencies

Trend-based functional dependencies model expectations on trends. They compare the behavior of two different parameters, which are related measurements within the same physiological context. For example, S_aO_2 and $P_{tc}O_2$ both give insight into the oxygenation of the patient. However, they react different in detail, but the global trend should be in parallel for both. We use the qualitative trend categories described in section 4.4 to compare the trends of such related parameters. The comparison is done using the short-term trend and the medium-term trend. If the trends differ by more than one category both measurements are marked as ambiguous.

A second usage of trend-based functional dependencies is to check whether the desired effect of a therapeutic action takes place. It is performed after a significant change of a parameter (ventilator setting), which controls the condition of the neonate. The method utilizes a specific delay time required to make a change in the ventilator setting visible in monitored parameters. For example, an increase of the inspired oxygen fraction F_iO_2 should cause an increase of the neonate's oxygen level O_2. This should be visible after a delay of 10 minutes in S_aO_2 and $P_{tc}O_2$.

The combination of inspecting trends of different parameters, which measure the same physiological context, with the inspection of trends after a therapeutic action gives a quite good insight into the validity of parameters. For example, if we find after an increase of FiO_2 that S_aO_2 is increasing, but $P_{tc}O_2$ is not, we can assume that $P_{tc}O_2$ giving invalid readings due to some other causes, like bad circulation.

3.3.4 Validation: Trend Assessment

The assessment of the parameter development examines the short-term trend. It compares two successive qualitative trend values of the parameter. An invalidity of the parameter is signaled if the trend categories are not the same or at least neighboring. The assessment procedure is applicable for the short-term trend only. The very-short-term trend reacts too rapidly to small oscillations of the values. The medium-term and the long-term trend are too insensitive.

The qualitative trend-categories are divided in an upper and a lower region by the normal region (see 4.4 for a detailed description of the process). According to these regions the ordering of the qualitative categories is defined as follows (compare figure 2):

- upper region: A1 - A2 - A3 - ZA - C
- lower region: B1 - B2 - B3 - ZB - D

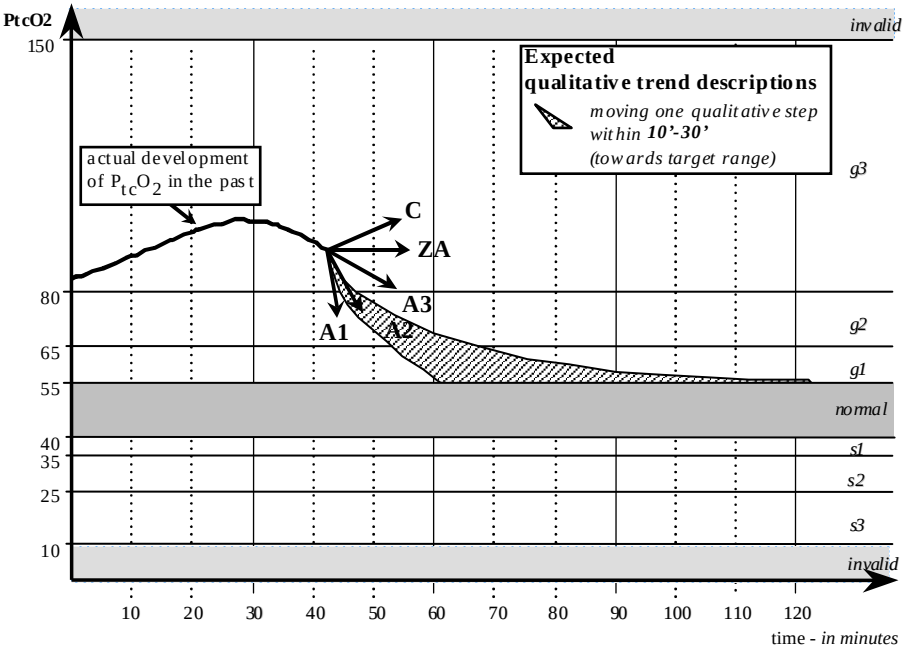

Figure 2: Trends defined in VIE-VENT. Values above the normal range can increase dangerously (C), stay constant (ZA), decrease too fast (A1), normal (A2), or too slow (A3).

The lower region is not shown in figure 2 but in figure 10 and in table 2.

The assessment procedure compares the previous qualitative short-term trend-category with the current one. If both belong to the same qualitative category or to a neighboring qualitative category then the parameter is validated as “correct”. Otherwise the parameter is classified as “wrong”.

The advantage of assessing qualitative trends is the ability to classify changes on a basis, which is better founded physiologically. For severe deviations from the target range we expect a return to the target range, which is fast initially and becomes slower and slower the nearer we approach the normal value. The trend-curve-fitting scheme and its resulting qualitative trend categories dynamically models this behavior.

3.3.5 Repair: Estimating Missing Values

During a monitoring process the position of a sensor has to be changed frequently and regularly. Therefore, the measurements are often missing. The implicit assumption of missing measurements during such a position change is that they will be steady keeping their previously observed values.

There are two possibilities to deal with missing measurements. First, a stepwise backward checking provides the last reliable value and we continue with this value as long as no other system change is detected. The reliability score of estimated values decreases over time and a user defined timeout prevents estimations based on values too old to be useful. Second, applying the growth rate of the short-term trend we estimate a "correct" value. A precondition is the stability of the trend. It is assumed to be true, if the medium-term and short-term qualitative trend-categories are identical. The trend-based estimation of a value is more accurate then simply using the last valid value, provided stability of the trend.

Estimating values is less problematic when the medical staff follows the general guideline that sensors should not be changed or calibrated during critical phases of the neonate. However, if we cannot get valid measurements over a longer period of time, the simplified reasoning process is applied (see section 3.1).

3.4 Time-Independent Validation: Reliability Ranking

This last category is based on time-independent reliability ranking of variables. From the medical and technical sampling point of view, there is a well-defined priority which measurement is more reliable than another, depending on different conditions.

This allows the definition of a reliability ranking scheme by the user. In case of contradicting parameter values which cannot be resolved by other methods, the more reliable one is selected according to the rating scheme.

Examples of reliability ranking of VIE-VENT are: arterial blood gases are more reliable than venous blood gases; invasive blood gases are more reliable than both transcutaneous blood gases and S_aO_2; S_aO_2 is more reliable than $P_{tc}O_2$. On the one hand these lists facilitate the data-validation task and on the other hand they also help the pruning of different and concurrent therapy recommendations.

4 Data Mining - Time-Oriented Data Abstraction

There are two fields of applications for the data abstracted from measuring devices in the clinical environment: Knowledge discovery and on-line monitoring.

Knowledge discovery retrospectively looks at recorded data to find significant patterns or to relate the raw data either to other information from

the patient data management system or to rules in a knowledge base (Fayyad & Uthurusamy, 1996). On-line monitoring (a.k.a. intelligent alarming) tries to detect dangerous situations in real-time and to suggest countermeasures to assist the physician in the treatment process.

Both applications share most of their requirements concerning the data abstraction process. The main difference lies in the point of view on the data—retrospective vs. real-time. In retrospective analysis all information for the whole period of interest is available. So for each point in time the past, the present and the future are known. Opposed to this, in real-time analysis, only past and present are known. In the following we will explain the general approach first and discuss the special aspects of on-line monitoring at the end of each subsection if appropriate.

We distinguish three basic qualitative abstractions of a curve at a given position: *state*, *grade*, and *bends*. *State* (section 4.1) is the qualitative expression of the value itself, e.g. slightly high, normal, or extremely low. *Grade* (section 4.2) is the first deviation or slope of the curve, e.g. slightly increasing or stable. Since both state and grade are not only extracted for a single instant, but for a interval of time, the output of the abstraction process is not a series of point data, but a sequence of time intervals, during which a certain qualitative value stays stable.

Bend (section 4.3) abstraction transforms the curve from a series of data points to a sequel of bends with lines in between. This representation matches the intuitive or naive terms users often use when describing curves like “first it goes up, then it makes a bend down, and then up again”.

Trend curve fitting (section 4.4) is a method to abstract trends as utilized in section 3.3 from state and grade of a value.

In section 4.5 we show how *derived status information* can be produced by rules in a knowledge base using basic qualitative abstractions and information from the validation process.

4.1 State Abstraction

The state of a value is its classification according to a list of qualitative values and their borders, also called qualitative regions. E.g. the qualitative region of normal transcutaneously measured pressure of CO_2 ($P_{tc}CO_2$) might be from 35 to 49 mmHg during ventilation mode IPPV. Figure 3 shows an example from VIE-VENT, where we defined seven qualitative regions: *s1*, *s2*, and *s3*, for increased values, *normal* for values within the target range and *g1*, *g2*, and *g3*, for decreased values.

The transformation from quantitative to qualitative values has four advantages: First, qualitative information is easier to comprehend than an number. Second, uniform rating schemes provide convenient access to the data for rules applied on that data. Third, equal rating schemes for different parameters make them comparable, even if their numerical values are not, and independent of the origin of data. Fourth, the maintenance of a knowledge base or reasoning component is facilitated if the medical knowledge about value of a qualitative regions' limits changed.

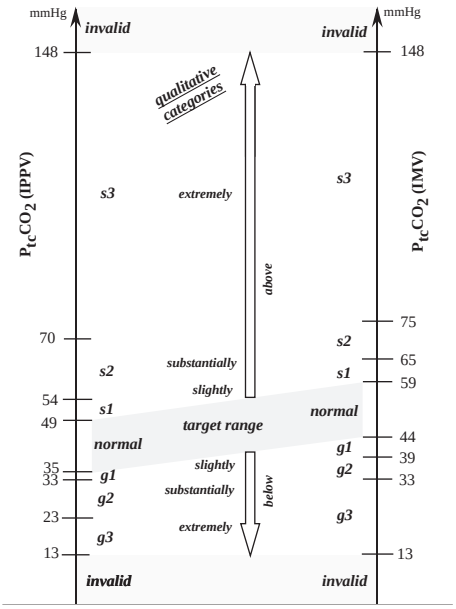

Figure 3: Qualitative regions of $P_{tc}CO_2$ in the context of IPPV and IMV.

Each set of qualitative regions is valid for a certain *context* only. The context is defined by the mode of ventilation—like in the above example—by the diagnoses, or by constraints like acute problems or defects of the patient. Changes in context are derived automatically from the input data.

Although this transformation seems simple, it is not when the data is noisy like the signals obtained from monitors in the ICU environment.

Sophisticated data validation (as described in section 3) can contribute a lot to the quality of high-frequency data. Still many signals recorded in the medical domain exhibit more or less small random oscillations, which are hard to separate from meaningful changes in the curve.

If such a curve would be changed to a series of intervals, during which the qualitative value stays stable, and if the curve oscillates on the border between two qualitative regions, these intervals would be too small to be meaningful. In some cases the range of oscillation can be wider than the width of each qualitative region which leads to unusable output for most of the recording time.

There are two simple remedies to the problem: Averaging and thresholds. Both fail, if the quality of the signal or the range of oscillation changes dynamically as it is the case in the medical domain (e.g. small movements of the patient lead to short periods of random oscillations in the measured S_aO_2). To cope with such cases, we developed a method to abstract qualitative values from a statistical representation of the distribution of data points at each part of the curve called spread. It is explain third.

4.1.1 Averaging

The first approach—averaging—simply means that each measurement is replaced by the average of a certain number of data points in its surrounding. The approach is quite simple but the number of points involved in the calculation and the kind of averaging (mean, moving average, etc.) are very sensible parameters. Too much smoothing (e.g. averaging too many data points) hides meaningful peaks in the curve while moderate smoothing still fails to suppress more significant oscillations.

4.1.2 Thresholds

The second approach—thresholds—defeats errors by imposing a threshold when crossing the border between two qualitative regions. Thus, the qualitative value only changes, if the quantitative value exceeds the borders of the current qualitative region by a certain percentage of its width—the threshold. To avoid excessive postponement of changes in cases where the quantitative value crosses the border of a qualitative region but does not exceed the threshold, a timeout period is defined, after which smoothing is terminated by defining the qualitative value according to the current quantitative value, even if it is near a region's border. Figure 4 illustrates such a smoothing algorithm implemented in VIE-VENT. The ε-region corresponds to the threshold and the activated period reflects the time period until the timeout is reached. In the example in figure 4, the smoothing takes place from time point t until $t + 7$, called the smoothing period. More details are given in (Miksch et al., 1993).

The problem lies in finding the best values for threshold and timeout. If they are too big, every change in qualitative value is unnecessary postponed. If it is too small, it does not suppress all undesired oscillations.

As with averaging good parameter settings might be found for curves of constant quality but no good solution can be found for dynamic changes in quality of measurement.

This algorithm was implemented in VIE-VENT (Miksch et al., 1993; 1996) but suffered from the inability to adjust the parameters in a way which fits the changing quality of the input.

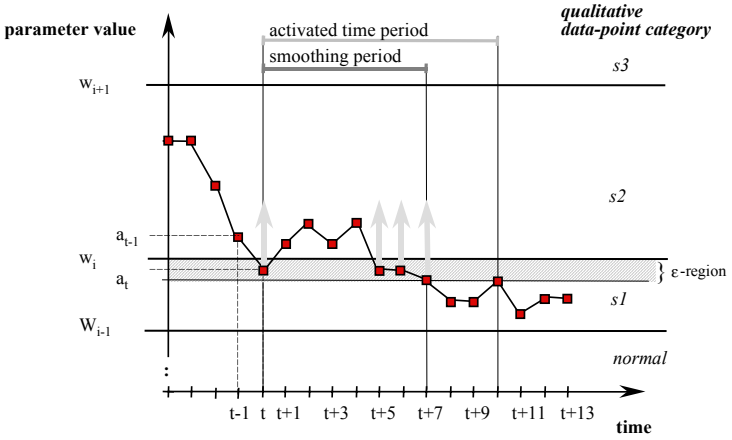

Figure 4: Smoothing of a curve oscillation on the border between two qualitative regions.

4.1.3 The Spread

To cope with these changing oscillations we developed a more complex representation of the curve, which we call the *spread* (Miksch et al., 1999). To calculate the spread, we slide a time window of constant width over the curve in small steps. For each position of the time window, we calculate a linear regression of the valid data points (i.e. not discarded by validation methods described in section 3) within the window. On the center of the line we plot the adapted standard error (s_a).

$$s_a = \frac{s}{\sqrt{N_{valid}}} \sqrt{N_{max}} \qquad (4)$$

This is the standard deviation (s) of the linear regression divided by the square root of the number of valid data points involved in the calculation

(N_{valid}) and multiplied by the square root of the maximum number of points possible in the interval (N_{max}).

Doing so we arrive at a vertical bar representing a statistically motivated estimation of the distribution of the data points in the time window. Connecting the ends of each bar with those of its neighbors yields a band (called *spread*), which vertically follows the average of the curve and the width of which shows the uncertainty involved in its calculation. The smaller the spread, the better the quality of the curve. Figure 5 shows the steps of the algorithm while figure 6 shows an example where the qualitative value is not influenced by a short peak which is not considered significant.

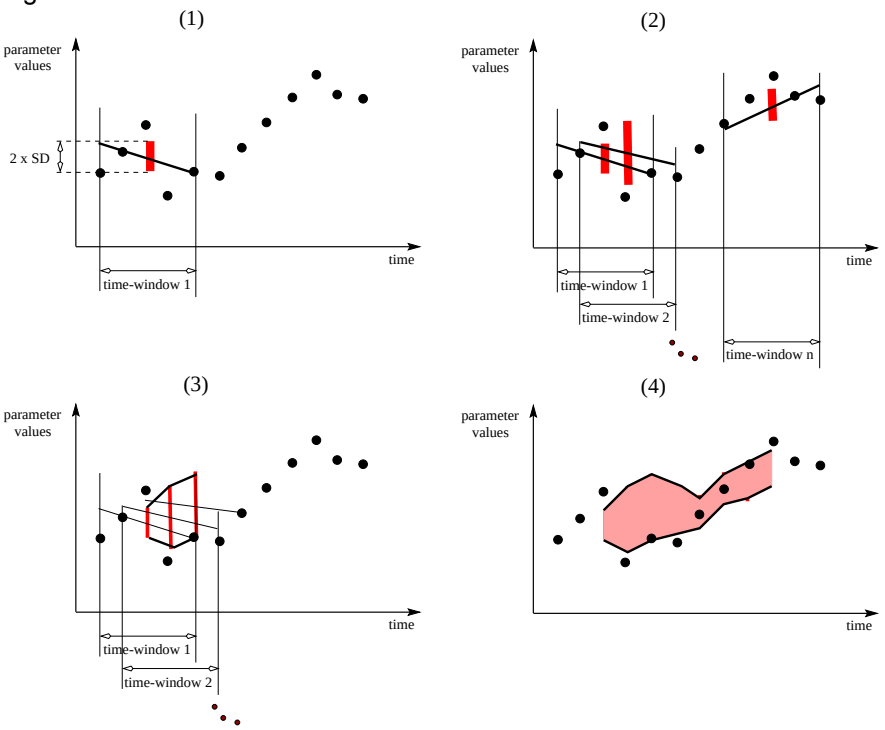

Figure 5: Calculation of the spread. For a given time-window (1) we calculate the linear regression line (long black line). On its center we plot the adapted standard error (equitation 4) up and down (gray vertical bar). This is repeated for all positions of the time-window (2). Connecting the ends of the bars (3) yields the spread (4). It represents the distribution of the data points based on statistical information.

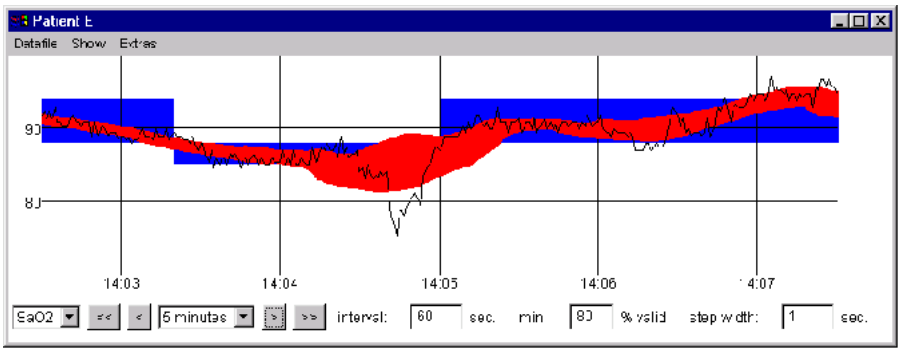

Figure 6: Practical example of the spread application. The thin line shows the raw data. The light gray area depicts the *spread*, the dark gray rectangles represent the derived temporal intervals of steady qualitative values. The lower part of the screen shot shows the parameters used. The length of the time window (interval) is set to 60 seconds. Thus the spread does not follow the short peak down at 14:04:45 but shows the deviation from the general trend by increased width.

The spread is used to abstract qualitative values of the curve. The qualitative value is changed only if both upper and lower margin of the spread are outside the previous qualitative region. This is a very conservative allocation strategy minimizing the changes in the qualitative value, but is mostly suitable for retrospective analysis.

For the purpose of intelligent alarming it might be desirable to plot the error bar on the rightmost end of the regression line instead of its center. In addition, alarms can also be triggered if only one margin of the spread crosses a certain limit e.g. if the upper margin of the spread enters the “extremely high”-region. Applying this procedure, bad data resulting in a wide spread cannot lead to delayed alarms, but might cause extraneous ones.

The advantage of the spread over other approaches lies in its dynamic adaptation to changing amounts of oscillations and missing data, which are very common in the clinical environment.

4.2 Grade Abstraction

While other authors like Shahar (Shahar & Musen, 1996) describe changes in the curve by two distinct qualitative values—gradient (e.g. increasing) and rate (e.g. fast)--we combine both in one value, the *grade*. It is the qualitative expression of the first deviation of the curve (e.g. fast increase, slow decrease) and can easily be derived from the slope of the regression lines calculated above. By drawing a spread for the slope too,

its advantages (as described above) can be used for the abstraction of the grade too.

While current monitors mostly rely on the measured value of the parameters—their states—there is a strong demand for systems doing a more sophisticated analysis of the measured data. Defining alarms for qualitative values of the grade in addition to those for the state can help to avoid critical situations by drawing the physicians attention to problems before they cause a crisis. To arrive at a meaningful picture the grades measured over different periods of time together with the state must be considered for each parameter.

4.3 Bend Abstraction

When asked to comment on a curve many people describe it as a series of lines with bends of different sharpness in between. Motivated by this, we developed an method (Miksch & Seyfang, 1999) to break a series of data points into a sequel of bends with lines connecting them. Bends in a curve can be detected by looking at the second deviation of its graph. A minimum there indicates a bend to the right or down on the original curve, a maximum indicates a bend to the left or up (see figure 7).

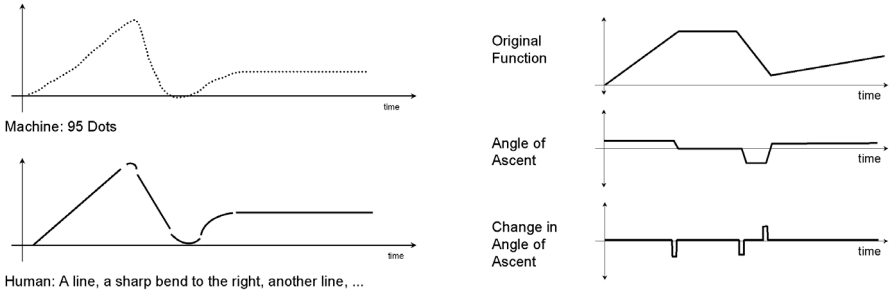

Figure 7: The bend abstraction. The basic idea in abstracting bends from a curve is, that humans describe as bends and lines what the devices supply as series of data points. Bends are defined as changes in the slope of the second deviation—in places, where the original curve makes a bend, its second deviation has an extremum.

To be more specific, first we calculate the angle of the slope of the original graph. Second, we calculate the slope of that curve. The resulting curve is called *indication function*. Each extremum or peek in the indication function represents one *bend*. From the sequel of the bends together with the original graph we derive *corners* and *lines*.

Each *bend* is described by the position of the corresponding extremum in the indication function, the height of the corresponding peak in the indication function, and the area of the peak.

The x-coordinate of the *corner* clearly equals the middle of the bend. The y-value could be the y-coordinate of the nearest point in the original curve, but to reduce influence of noise, it is necessary to take the average of some of its neighbors into account too in most cases. Integrating too many of them in the calculation will distort the result towards the inner side of the bend.

The *lines* between the bends can either be drawn just as connections of the corners of the curve, or they are calculated as a linear regression of the points of the original curve between two bends.

Which version of the above definitions is taken depends on the focus or preference of the users, which varies between different domains of application.

The data abstracted this way can be used in three different ways:

4.3.1 Direct Visualization

An example for the direct visualization of bends is shown in figure 8. Each bend in the original graph (at the top) is represented by a bar (on the bottom), who's height and area equals the height and area of the corresponding peak in the indication function (in the middle). This method is applicable to tasks, where the attention of the user must be drawn to relatively small irregularities in a periodic curve.

4.3.2 Symbolic Representation

The information about the bend can be expressed in list to make in accessible to symbolic reasoners like knowledge based systems or machine learning tools. This is important to bridge the gap between raw data delivered by monitors and knowledge bases using this knowledge in a high-level way.

The following Example describes a graph consisting of a line increasing by 20 degrees for 100 seconds followed by a narrow bend to the right and 30 seconds of decrease.

```
((line 100sec up 20°)
 (bow right narrow)
 (line 30sec down 30°))
```

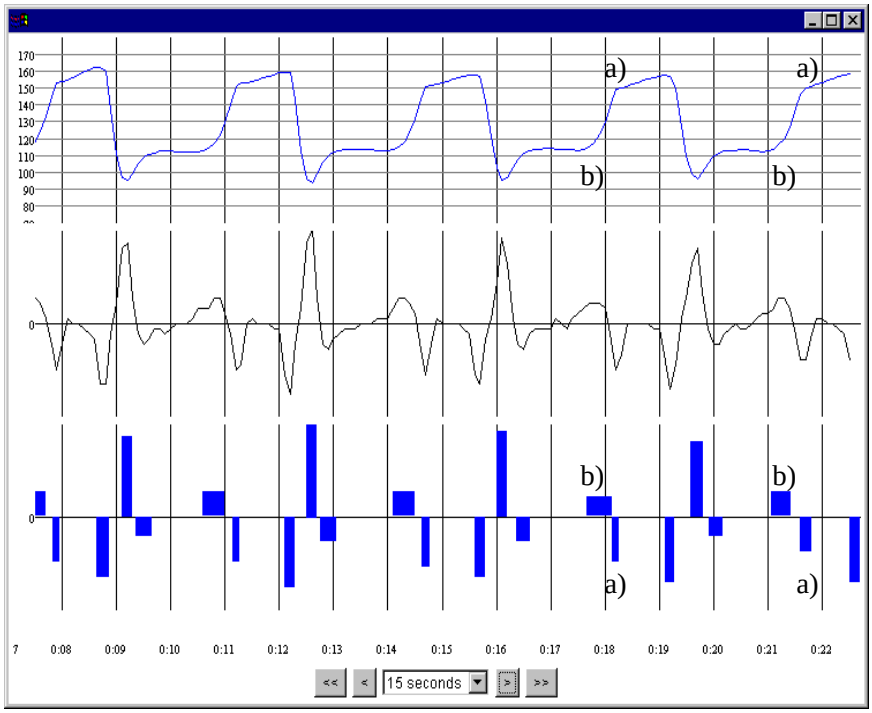

Figure 8: Visualization of bends. Starting at the top, we show the original graph, the indication function and the bars representing the significant bends. a) shows an example of an irregularity in the curve which a human could also detect if concentrating on every detail: the bend in the right-most oscillation is not as sharp as the corresponding one in the other oscillations. b) draws our attention to a feature not perceptible by looking at the raw data: the long-spread bow to the left of the second oscillation from the right is not as sharp as the others as indicated by inferior height of the bar. The corresponding part of the original curve does not seem different by itself. The significance of the features found in a) and b) depends on the domain knowledge about the data represented by the curve.

4.3.3 Interoscillation Reasoning

Many types of data recorded in the medical domain are periodic but varying. Deviations have natural, pathological, technical, or unexplainable origin. The field of signal processing provides a wide range of methods for both noise reduction and detection of deviations. However, they are designed for signals with technical origin, for which an exact mathematical model is available. Many signals recorded in medicine lack such an exact model. In some cases, there is a qualitative model, roughly describing the interdependencies of some parameters, but not supplying an exact formula for the calculation of the “real” values.

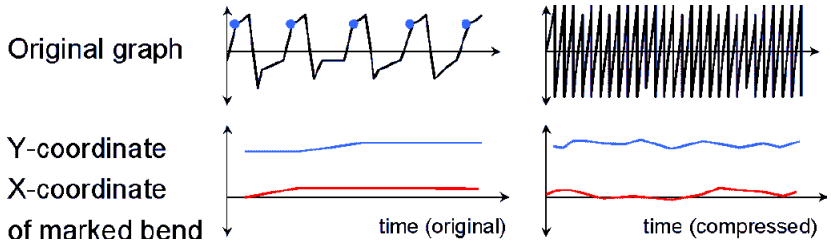

Figure 9: Visualizing the changes of the oscillation's shape. Plotting a bend's X-coordinate (its offset within the oscillation) or the Y-coordinate (the value of the original curve at this position) on a separate graph yields a very dense representation of the bend indicating all its changes and deviations clearly.

Another aspect is the format, in which the curve or function is described. Fast Fourier analysis, for example, describes periodic curves by a list of frequencies and their amplitudes. For a physician this is a very unusual way to look at an ECG.

As an alternative, we compare the position of each bend in an oscillation to its position in other oscillations. This yields information on the change of the oscillation's appearance over time which is rather intuitive, since it is in terms like "the second peak moved up by 10 % over the last minute". In many fields of application such formulations are compatible with the knowledge acquired by human experts when looking at curves (without the aid of computer systems).

On the one hand, plotting one dimension of a corner point in each instance of an oscillation as a separate curve yield a graph which gives a clear picture of the corner's development over time, even when the time axis is compressed (figure 9). On the other hand, this graph can be used as input to data abstraction to obtain qualitative information about the development of that detail of the oscillation.

As of this writing, we are examining the applicability of this approach to the fields of ECG and CTG.

4.4 Trend Curve Fitting

Often the aim of a treatment correlates with bringing the value of some parameter back to the normal region. In these cases the grade alone does not give full information. Instead we are interested in learning whether the value is improving i.e. approaching the normal region or not. An increase of a value which is too high has different semantics than an increase of a value which is too low.

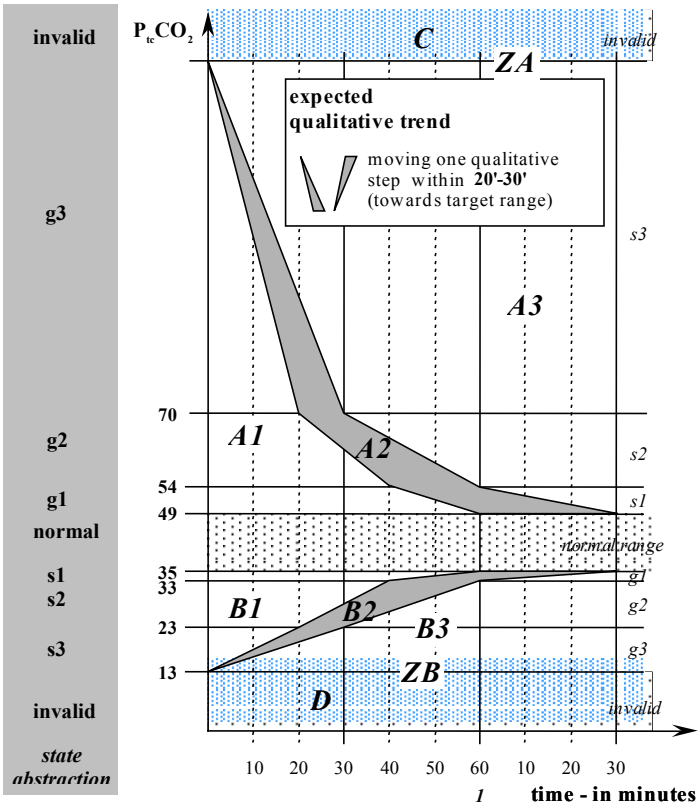

Figure 10: Expected improvement $P_{tc}CO_2$. Values which are too low should improves by one qualitative step within 10 to 20 minutes while values which are too high may take 20 to 30 minutes per step to improve.

In many cases there is also an estimation of how long it takes under normal conditions for the value to return to normal. So we can distinguish between normal improvement, too slow improvement and too fast improvement. Figure 10 shows the trends implemented in VIE-VENT. In this example, we demand that the value of $P_{tc}CO_2$ improves by one qualitative step within 20 to 30 minutes. The algorithm is explained in (Miksch et al., 1993). Table 2 summarizes the trends of a curve, depending on its state and rate of change.

values above the target range are	
A1	decreasing too fast
A2	decreasing normally
A3	decreasing too slow
ZA	staying constant
C	increasing dangerously

values below the target range are	
B1	increasing too fast
B2	increasing normally
B3	increasing too slow
ZB	staying constant
D	decreasing dangerously

Table 2: Schema of trend curve fitting in VIE-VENT.

4.5 Derived Status Information

In many cases the trend or state of a parameter does not itself give enough information. A rule base is needed to abstract more useful information from the basics.

The qualitative temporal abstraction of monitoring variables makes it easy to use simple rules to activate therapeutic actions. For example, rule *R8-therapeutic-actions* states, that we recommend an increase of both frequency and *PIP*, if the short-term trend of $P_{tc}CO_2$ is *A3*, *ZA*, or *C*, and its state is above normal, i.e. *s1*, *s2*, or *s3*, and the input is classified as correct.

```
(defrule R8-therapeutic-actions
activate-therapeutic-action-PtcCO2-ventilation"
   (phase (kind therapy_recommendation))
   (ventilation_phase (kind ippv))
   ?f1 <- (thp_recommendation ventilation)
   (qual_trend_category (parameter PtcCO2)
                        (kind_of_trend short)
                        (qual_trend A3|ZA|C))
   (qual_data_point_category (parameter PtcCO2)
                             (site tc)
                             (value s1, s2, s3))
   (causal-explanation-validation (parameter PtcCO2)
                                  (classification correct))
=>
   (retract ?f1)
   (assert (action (reason ventilation)
                   (BG PtcCO2)
                   (kind inc-f))
           (action (reason ventilation)
                   (BG PtcCO2)
                   (kind inc-pip))))
```

The essential preconditions for triggering therapeutic actions depend on the qualitative trend abstraction of the short-term trend (expressed as

qual_trend_category in the rule *R8-therapeutic-actions*) and the qualitative state abstraction (expressed as *qual_data_point_category* in the rule *R8-therapeutic-actions*). If the qualitative state abstraction is *s1* or *s2* or *s3*, and the qualitative trend abstraction is *A3* or *ZA* or *C*, then therapeutic actions are recommended (increase ventilator settings). The second fact *ventilation_phase* in the left-hand side (LHS) of rule *R8-therapeutic-actions* refers to the mode of ventilation (i.e., IPPV) and indicates, that this rule belongs to the set of rules dealing with the phase of Intermittent Positive Pressure Ventilation). The last fact, *causal-explanation-validation*, supplies the necessary explanations of the data validation process, namely the classification of the particular validated parameter. The right-hand side (RHS) of rule *R8-therapeutic-actions* specifies the therapeutic actions. Each action-fact includes the kind of the recommended action and an explanation of the circumstances: the fact *(reason ventilation)* refers to "ventilation" process depending on the system model of ventilation), *(BG PtcCO2)* refers to the relevant parameter, namely the blood gas measurement, and *(kind ?x)* determines which particular action has to take place (e.g. *(kind inc-pip)* means, that an increase of the peak inspiratory pressure (*PIP*) is recommended).

5 Evaluation and Discussion

5.1 Empirical Evaluation

Within the VIE-VENT system (Miksch et al., 1993; 1996), we evaluated the effectiveness of the above data-validation methods presented in section 3 utilizing a particular evaluation scenario consisting of two steps (Horn et al., 1997): first, a visual inspection of the results of the data validation process, and, second, a formal evaluation of the validation results.

Our sample consists of 640786 seconds (approx. 177 hours) of data recordings from nine neonates. The age of the neonates was between four days and six weeks, the weight between 690 g and 3460 g.

In the first step, the data from the first six patients (approx. 115 hours) were used to tune the validation parameters, specifically to find suitable parameters for the stability check and the Højstrup method. Additional validation parameters, which could not be determined from the data recordings, were derived from the knowledge of expert neonatologists. We plotted the data curves and annotated the invalid data with rectangular markers below each curve, when our data-validation methods recognized errors. Two expert neonatologists examined the results. The parameters in

our data-validation methods were tuned towards the overall goal of avoiding wrong therapeutic recommendations. As a consequence, the data-validation methods marked all measurements as invalid which depicted *unusable* signals. The remaining data (approx. 62 hours of recording) have been used to verify the correctness of the data-validation methods.

In the second step in our evaluation scenario we compared the data points found invalid by an expert neonatologist with the invalidation markers produced by VIE-VENT. Currently, no widely accepted "gold standard" exists to judge the correctness of the continuously assessed data of $P_{tc}O_2$, $P_{tc}CO_2$ and S_aO_2. Therefore, we relied on the judgement of the domain experts, experienced neonatologists.

For this evaluation study we took sequences of continuously assessed data which show some variation. We selected continuous sequences of 4320 seconds length which contain at least two invalidation markers from VIE-VENT.

From these sequences we randomly selected five sequences from different patients. The selected sequences were presented to the expert using high resolution plotting (without the invalidation markers of VIE-VENT). The expert marked those data points which he judged invalid. Table 3 gives the evaluation results from the comparison of the expert's and VIE-VENT's invalidation markers. VIE-VENT's perfect sensitivity is not surprising due to the tuning of the parameters towards recognition of all artifacts and unclear trends. The rather low specificity results from the overall goal to avoid wrong therapeutic recommendations. A further complication which lowers specificity is the fact, that the expert is able to see the future development of a parameter from the plot. In contrast, VIE-VENT operates in real-time. It has to wait for stability of a parameter until it is set back valid. This increases the number of false positives but is an effect caused by the constraints of real-time operation.

Parameter	*Sensitivity*	*Specificity*
S_aO_2	100%	88.9%
$P_{tc}O_2$	100%	83.2%
$P_{tc}CO_2$	100%	94.6%

Table 3: Evaluation of VIE-VENT's data validation procedures.

5.2 Discussion

The high-level abstraction methods presented in section 4 lead to the a series of opportunities both in visualisation and in interfacing knowledge-based systems.

5.2.1 Compact Visualization

Displaying only the important features of a graph in an abstract form in addition to the original graph allows for easy detection of trends and outliers which otherwise would be burried in the overwhelming impression of the data. Currently we are investigating into the application of various abstract data representations in the field of ECG-analysis and ventilation monitoring.

5.2.2 Bridge to Knowledge Representation

The abstracted information can be matched against conditions in a rule base. So the curves can be tagged according to a set of classifications stored in a knowledge base. This aspect is crucial for the integration of high-frequency data and symbolic systems such as symbolic machine learning, knowledge-based systems for intelligent alarming and a guideline execution system like the one developed in the Asgaard (Shahar, Miksch, & Johnson, 1998) project.

5.3 Overall Benefits

In the following we are summarizing the main benefits of our proposed methods:

5.3.1 Improving the Quality of Data

The data recordings assessed from various channels of devices are more erroneous than commonly expected. Applying our validation methods to the observed on-line and off-line data sets resulted in automatic elimination of most invalid measurements: false positive alarms were reduced and errors of data interpretation were minimized.

5.3.2 Communicating Various Kinds of Time-Stamped Data Lucidly

The physicians need an overview over a certain period of time and over various parameters which together give a more detailed, reliable, and comprehensible picture of the patient’s condition. Our time-oriented data-

abstraction methods transform a huge amount of numerical, time-stamped values into a convenient set of easy to understand qualitative descriptions of the patient's situation. This results in diminishing the information overload by visualizing the available information in a user specific and capable way: the physicians can recognize and predict a critical patient's condition more easily, which finally ensures a better treatment management.

5.3.3 Bridging to Higher-Level Reasoning

More sophisticated reasoning tasks need more advanced representations than numerical data or simple qualitative assessments. Our approach facilitates tagging various time-oriented data sets according to a set of qualitative and more intuitive classifications. These qualitative characteristics over time can be matched against conditions from a rule base, which results in more obvious and simple rule base. At the same time, this rule base enables more powerful reasoning components to be applied.

6 Future Research

Future work will focus on extending the methods' understanding of the underlying processes and on running elaborate evaluations.

6.1 Utilizing Qualitative Descriptions in Treatment Planning

Previous version of VIE-VENT used standard forward-chaining rules to formulate the knowledge about the data and its interdependencies. But application domains like artificial ventilation of neonates can only be described fully as a set of interweaving and interdependent treatment processes.

Several researches try to formalize that in knowledge-based systems. Still, for domains like ventilation in ICUs, most approaches do not seem powerful enough (Miksch, 1999). Instead, a framework for time-oriented modeling of treatment procedures (Shahar, Miksch, & Johnson, 1998) is needed to proper represent the domain knowledge, which can lead to even higher level abstraction of the data such as the assistance of the weaning process through recommendations for the settings of the respirator.

6.2 Repetitive Temporal Patterns

In dynamically changing environments, like ICUs, the basic temporal data-abstraction methods resulting in qualitative state, gradient, rate, or simple pattern description (compare (Shahar & Musen, 1996; Shahar 1997; Miksch et al., 1996)), are not sufficient.

High-level temporal data-abstraction methods are needed, which include a wide variability in the behavior of a parameter, variability in the time patterns a parameter shows, and specifications for relating different temporal patterns of different parameters. They have to be able to recognize and to describe recurring states, events, episodes, or actions (e.g. information about the frequency of temporal patterns in the past, like "three episodes of hyperoxemia during the last three hours occurred").

6.3 Further Evaluation

To achieve acceptance by practitioners it is crucial to run extensive evaluations both on recorded data and in real-time with all parts of the system. Currently, we are working on evaluation scenarios to examine the usefulness of our approach in the clinical setting of artificial ventilation of newborn infants.

6.4 Conclusion

We described methods to validate and to repair potentially unreliable time-oriented, high-frequency data and to abstract different kinds of qualitative descriptions over time from the validated but still partially untrustworthy data, in which some artifacts might remain unrecognized. Our methods presented were successful in overcoming the problems the medical staff is facing currently: first, to improve the quality of data to arrive at trustworthy data and, second, to ease the information overload caused by various channels of on-line and off-line data recordings. The methods support the medical staff to easily comprehend the various continuously and discontinuously assessed data utilizing different qualitative abstractions over time and combination thereof.

Our approach was evaluated on data from artificial ventilation of neonates and proved to reduce the number of wrong alarms while correcting most artifacts in the data.

Future research will focus on adding high-level treatment planning to the domain-knowledge base to implement an even deeper understanding of the processes lying behind the observed data. This will improve the

performance especially in non-standard cases like life-threatening situations which cannot be described by a simple set of rules.

Acknowledgments

We thank Georg Duftschmid, Gerhilde Egghart, Klaus Hammermüller, and Robert Kosara for their contributions. This project is supported by "Fonds zur Förderung der wissenschaftlichen Forschung - FWF" (Austrian Science Foundation), P12797-INF. We greatly appreciate the support given to the Austrian Research Institute of Artificial Intelligence (OFAI) by the Austrian Federal Ministry of Science and Transport, Vienna.

References

Avent, R., and Charlton, J. 1990. A critical review of trend-detection methologies for biomedical monitoring systems. *Critical Reviews in Biomedical Engineering* 17(6):621-659.

Bellazzi, R.; Larizza, C.; Magni, P.; Montani, S.; and De Nicolao, G. 1999. Intelligent analysis of clinical time series by combining structural filtering and temporal abstractions. In Horn, W.; Shahar, Y.; Lindberg, G.; Andreassen, S.; and Wyatt, J., eds., *Proceedings of the Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making, AIMDM'99*, 271-280. Berlin: Springer.

Carlson, D.; Wallace, J.; East, T.; and Morris, A. 1993. Verification and validation algorithms for data used in critical care decision support systems. In Gardner, R., ed., *Proceedings 19th Annual Symposium on Computer Applications in Medical Care (SCAMC'95)*. 188-192.

Cios, K.; Pedrycz, W.; and Swiniarski, R. 1998. *Data Mining Methods for Knowledge Discovery*. Boston: Kluwer Academic Publishers.

Clancey, W. J. 1985. Heuristic classification. *Artificial Intelligence* 27:289-350.

Egghart, G. 1995. Validierung von In- und Outputdaten: Methoden und ihre Anwendungen in VIE-VENT. Master's thesis, Department of Medical Cybernetics and Artificial Intelligence, University of Vienna.

Fayyad, R., and Uthurusamy, R. 1996. Data mining and knowledge discovery in databases. *Communications of the ACM* 39:24-26.

Gardner, R.; Hawley, W.; East, T.; Oniki, T.; and Young, H.-F. 1992. Real time data acquisition: Experience with the medical information bus (MIB). In Clayton, P., ed., *Proceedings of the Fifteenth Annual Symposium on Computer Applications in Medical Care, SCAMC-91*. New York: McGraw-Hill.

Garfinkel, D.; Matsiras, P.; Mavrides, T.; McAdems, J.; and Aukburg, S. 1989. Patient monitoring in the operating room: Validation of instruments reading by artificial

intelligence methods. In Kingsland, L., ed., *Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care, SCAMC-89*, 575-579. Washington, DC: IEEE Computer Society Press.

Haimowitz, I.; Le, P.; and Kohane, I. 1995. Clinical monitoring using regression-based trend templates. *Artificial Intelligence in Medicine* 7(6):473-496.

Højstrup, J. 1992. A statistical data screening procedure. *Measurement Science & Technologie* 4:153-157.

Horn, W.; Miksch, S.; Egghart, G.; Popow, C.; and Paky, F. 1997. Effective data validation of high-frequency data: Time-point-, time-interval-, and trend-based methods. *Computer in Biology and Medicine, Special Issue: Time-Oriented Systems in Medicine* 27(5):389-409.

Hunter, J., and McIntosh, N. 1999. Knowledge-based event detection in complex time series data. In Horn, W.; Shahar, Y.; Lindberg, G.; Andreassen, S.; and Wyatt, J., eds., *Proceedings of the Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making, AIMDM'99*, 271-280. Berlin: Springer.

Kay, S. 1993. *Fundamentals of Statistical Signal Processing*. Engelwood, New Jersey: Prentice Hall.

Keravnou, E. T. 1997. Temporal abstraction of medical data: Deriving periodicity. In Lavrac, N.; Keravnou, E.; and Zupan, B., eds., *Intelligent Data Analysis in Medicine and Pharmacology*. Boston: Kluwer Academic Publisher. 61-79.

Larizza, C.; Bellazzi, R.; and Riva, A. 1997. Temporal abstractions for diabetic patients management. In *Proceedings of the Artificial Intelligence in Medicine, 6th Conference on Artificial Intelligence in Medicine Europe (AIME-97)*, 3 19-30. Berlin: Springer.

Lawless, S. 1994. Crying wolf: False alarms in a pediatric intensive care unit. *Critical Care Medicine* 22:981-985.

Miksch, S., and Seyfang, A. 1999. Finding intuitive abstractions of high-frequency data. In Miksch, S.; Andreassen, S.; Dojat, M.; Hunter, J.; Popow, C.; Rees, S.; and Thorgard, P., eds., *Workshop "Computers in Anaesthesia and Intensive Care: Knowledge-Based Information Management", in conjunction with the Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making (AIMDM'99)*, 13-26.

Miksch, S.; Horn, W.; Popow, C.; and Paky, F. 1993. VIE-VENT: Knowledge-based monitoring and therapy planning of the artificial ventilation of newborn infants. In Andreassen, et al., S., ed., *Artificial Intelligence in Medicine (Proceedings of the Fourth European Conference on Artificial Intelligence in Medicine Europe (AIME-93)*. Amsterdam: IOS Press.

Miksch, S.; Horn, W.; Popow, C.; and Paky, F. 1996. Utilizing temporal data abstraction for data validation and therapy planning for artificially ventilated newborn infants. *Artificial Intelligence in Medicine* 8(6):543-576.

Miksch, S.; Seyfang, A.; Horn, W.; and C., P. 1999. Abstracting steady qualitative descriptions over time from noisy, high-frequency data. In Horn, W.; Shahar, Y.; Lindberg, G.; Andreassen, S.; and Wyatt, J., eds., *Proceedings of the Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making, AIMDM'99*, 281-290. Berlin: Springer.

Miksch, S. 1999. Plan management in the medical domain. *AI Communications*, 4.

Quinlan, J. 1986. The effect of noise on concept learning. In Michalski, R.; Carbonell, J.; and Mitchell, T., eds., *Machine Learning. An Artificial Intelligence Approach*. Boston: Morgan Kauffmann. 149-166.

Shahar, Y., and Musen, M. A. 1996. Knowledge-based temporal abstraction in clinical domains. *Artificial Intelligence in Medicine, Special Issue Temporal Reasoning in Medicine* 8(3):267-98.

Shahar, Y.; Miksch, S.; and Johnson, P. 1998. The Asgaard Project: A task-specific framework for the application and critiquing of time-oriented clinical guidelines. *Artificial Intelligence in Medicine* 14:29-51.

Shahar, Y. 1997. A framework for knowledge-based temporal abstraction. *Artificial Intelligence* 90(1-2):267-98.

Sittig, D., and Factor, M. 1990. Physiologic trend detection and artificial rejection: a parallel implementation of multi-state kalman filtering algorithm. *Computer Methods and Programs in Biomedicine* 31:1-10.

Sittig, D., and Orr, J. 1992. Evaluation of a parallel implementation of the learning portion of the backward error propagation neural network: Experiments in artifact identification. In Clayton, P., ed., *Proceedings of the Fifteenth Annual Symposium on Computer Applications in Medical Care, SCAMC-91*, 290-294. New York: McGraw-Hill.

Sukuvaara, T.; Koski, E.; Maekivirta, A.; and Kari, A. 1992. A knowledge-based alarm system for monitoring cardiac operated patients - technical construction and evaluation. *International Journal of Clinical Monitoring and Computing* 10:117-126.

Uckun, S. 1994. Intelligent systems in patient monitoring and therapy management. *International Journal of Clinical Monitoring and Computing* 11:241-253.

van der Aa, J. 1990. Intelligent alarms in anesthesia. Master's thesis, Technische Universiteit Eindhoven, The Netherlands.