Effective Data Validation of High-Frequency Data: Time-Point-, Time-Interval-, and Trend-Based Methods

Werner Horn [a,b], Silvia Miksch [a,1], Gerhilde Egghart [b], Christian Popow [c], and Franz Paky [d]

[a] *Austrian Research Institute for Artificial Intelligence*
Schottengasse 3, A-1010 Vienna, Austria
Email: werner@ai.univie.ac.at

[b] *Department of Medical Cybernetics and Artificial Intelligence, University of Vienna*

[c] *NICU, Division of Neonatology, Department of Pediatrics, University of Vienna*

[d] *Department of Pediatrics, Hospital of Mödling, Austria*

Abstract

Real-time systems for monitoring and therapy planning, which receive their data from on-line monitoring equipment and computer-based patient records, require reliable data. Data validation has to utilize and combine a set of fast methods to detect, eliminate, and repair faulty data, which may lead to life-threatening conclusions. The strength of data validation results from the combination of numerical and knowledge-based methods applied to both continuously-assessed high-frequency data and discontinuously-assessed data.

Dealing with high-frequency data, examining single measurements is not sufficient. It is essential to take into account the behavior of parameters over time. We present time-point-, time-interval-, and trend-based methods for validation and repair. These are complemented by time-independent methods for determining an overall reliability of measurements. The data validation benefits from the temporal data-abstraction process, which provides automatically derived qualitative values and patterns. The temporal abstraction is oriented on a context-sensitive and expectation-guided principle. Additional knowledge derived from domain experts forms an essential part for all of these methods.

The methods are applied in the field of artificial ventilation of newborn infants. Examples from the real-time monitoring and therapy-planning system VIE-VENT illustrate the usefulness and effectiveness of the methods.

Keywords: data validation, temporal reasoning, high-frequency domains, real-time systems in medicine, ICU.

1 Introduction

Intensive care units (ICUs) are well equipped with most 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 a regular activity in today's ICUs. Even in the early years of ICU's data acquisition it was quite clear that patient data must be as complete as possible and that stored data should be free of artifact [1]. Today, monitors have builtin alerts, but the result is a vast volume of false alarms [2]. Alarming systems based on simple range checks are obviously too simple to be useful in a complex medical setting.

In the last years, several sophisticated knowledge-based monitoring and therapy-planning systems have been introduced [3]. These systems concentrated on optimizing data analyses and interpretation based on temporal abstraction mechanisms, on applying different kinds of accessible knowledge and information to enrich the reasoning process, and on minimizing manual data input as a result of the improvement of technical equipment at modern clinics and of access to computer-based patient records. Their usefulness will increase extremely when used as intelligent real-time control systems integrating the results of many sensor readings coming at various rates from a patient and presenting a whole picture [4]. Nevertheless, such monitoring and therapy-planning system to become effective and efficient requires reliable data [5]. Data received from monitors are more faulty than is often realized. Ten years after his request for artifact-free data, Gardner et al.[6] still reports about inspired oxygen fraction (FiO_2) recordings being correct only about 50% of the time. The importance of data validation has been neglected in the past. Real-time systems in medicine will not become operational without intensive efforts to detect artifacts. This requires combining all information available, cross-validating data sources, inspecting and reasoning about data points, and looking at trends to get a complete and consistent picture of the situation of the patient.

In the following section we will discuss the need for effective data validation. The approach taken in our monitoring and therapy-planning system VIE-VENT is shown in section 3. Methods for data abstraction and data validation are presented in detail in sections 4 and 5. In section 6 we discuss these methods. The results of the evaluation are shown in section 7. Section 8 summarizes the work and its conclusions.

[1] currently visiting scholar at Knowledge Systems Laboratory, Stanford University.

2 The Need for Effective Data Validation

We evaluated on-line data sets obtained from newborn infants with various respiratory illnesses. The data were collected from the monitoring system of a neonatal intensive care unit (NICU) once per second (between 16-28 hours continuous data recording for each newborn infant). The data sets consist of measurements of the transcutaneous partial pressure of oxygen ($P_{tc}O_2$) and carbon dioxide ($P_{tc}CO_2$), the heart rate (HR) given from ECG, the oxygen saturation (S_aO_2), and the pulse frequency ($PULS$) given from pulsoximetry. We combined these data sets with additional off-line data acquired from the CPR. Off-line data include ventilator settings (PIP, $PEEP$, F_iO_2, frequency f, etc.), results of invasive blood-gas analyses (pH, P_aO_2, P_aCO_2, where a denotes a measurement from arterial blood—we have venous and capillary measurements too), and clinical parameters (e.g., spontaneous breathing effort).

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 trustable data for future reasoning. First, 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. Second, 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). Third, 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. Fourth, additional invalid measurements were caused by on-line transmission problems or were unexplainable.

Up-to-now, data validation concentrated on numerical methods. These methods are successful for particular problem characteristics detecting values, which are not within certain ranges and trend values, which are physiologically implausible. At least range checking facilities are standard for today's monitors in ICUs. However, they result in numerous false alarms—or, if switched off—missing alarms [2]. Most of these numerical methods do not allow to 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. They give 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 values and trend schemata.

3 Data Validation in VIE-VENT

VIE-VENT [7–9] is an open-loop, knowledge-based monitoring and therapy-planning system for artificially-ventilated newborn infants. Our aim in developing VIE-VENT was to incorporate alarming, monitoring, and therapy-planning tasks within one system in order to overcome the limitations of existing systems, like GUARDIAN [10], SIMON [11,12], and NeoGanesh [13,14]. VIE-VENT is especially designed for practical use under real-time constraints at neonatal ICUs. Its various components are built in analogy to the clinical reasoning process. The data-driven architecture of VIE-VENT consists of several modules: data selection, data validation, data abstraction, data interpretation and therapy planning. All these steps are involved in a single cycle of data collection from monitors. The data selection module filters out context-relevant data for further processing. Data validation and data abstraction are discussed within this paper. Data interpretation classifies the state of the respiratory system of the newborn infant based on the unified qualitative parameters received from the data abstraction module. The therapy-planning module formulates therapeutic actions based on the interpretation of monitoring data, prunes therapeutic actions, and verifies whether the actions are effective. VIE-VENT's system model represents the neonatal respiratory function by two processes: ventilation (CO_2 elimination) and oxygenation (oxygen uptake). The output of the system are mainly recommendations for changing respirator settings. Additionally, VIE-VENT issues warnings in critical situations, as well as comments and explanations about the state of the respiratory system of the newborn infant.

During three years of development and evaluation of VIE-VENT we have learned that high-frequency data received from monitors in ICUs are not that accurate one would expect from modern equipment. Especially, non-invasive on-line acquired measurements result in data, which are rather vague. These measurements depend on the correct placement of sensors, the circulation of the neonate, body movements, and environmental conditions. Even regular sensor application and calibration may cause deviations and errors. Thus, it becomes less interesting to interpret the exact values but rather get reliable answers to three questions:

(i) Is the reading valid?
(ii) How is the reading to be qualified, e.g., being normal or substantially deviated from the normal range?
(iii) Are we able to qualify the trend, e.g., leading towards the normal range

or dangerously deviating from it?

In order to answer question 2 a transformation of quantitative data points into qualitative values is needed. The derived qualitative values form the basis for transforming interval data into qualitative trend descriptions, which should answer question 3.

To answer question 1 we performed a data-validation process based on various kinds of real-time data (high and/or low frequency, continuously- and/or discontinuously-assessed, and quantitative and/or qualitative data) and on different temporal ontologies (time points, time intervals, and trends): first, context-sensitive examination of the plausibility of input data and second, applying repair and adjustment methods for correcting erroneous or ambiguous data. To classify the input data we combined and enhanced established techniques (e.g., causal and functional dependencies) with newer techniques, based on qualitative descriptions, during different time periods.

Data abstraction is discussed prior to the presentation of the data-validation methods being a prerequisite for doing data validation based on qualitative values. However, data abstraction and data validation are two processes which are strongly intertwined. It is not possible to compute qualitative trends without having reliable data for a majority of time points included in the trend. On the other hand having evaluated a trend it may become clear that the last data-point value is erroneous. The interaction of the methods is discussed in section 5.5.

4 Temporal Data Abstraction

The usage of qualitative descriptions for data validation requires a temporal data-abstraction process, which derives qualitative values from the numerical data values received. The aim of the data abstraction process is to arrive at unified qualitative patterns for all parameters. The temporal abstraction methods are context-sensitive and expectation-guided. They incorporate knowledge about data points, time intervals and expected qualitative trend descriptions. In addition to its usage for data validation, derived qualitative values constitute the essential basis for the data interpretation and therapy-planning phase of VIE-VENT.

The transformation of data points is discussed in the next subsection, followed by a section which presents the trend curve fitting schemata. Having derived qualitative data-point values and qualitative trends we proceed with the data validation task.

Table 1
The unified scheme for abstracting blood-gas measurements.

Code	Category	
g3	extremely	below
g2	substantially	
g1	slightly	
normal	target range	
s1	slightly	above
s2	substantially	
s3	extremely	

4.1 An Unified Scheme for Data-Point Transformation

The transformation of quantitative data into qualitative values is usually performed by dividing the numerical range of a parameter into regions of interest. Each region represents a qualitative value. The region defines the only common property of the numerical and qualitative values within a particular context and at a specific time-stamp. It is comparable to the "point temporal abstraction" task of [15,16].

VIE-VENT uses context-sensitive schemata for data-point transformation of blood-gas measurements. The result of the abstraction process is an unified scheme for all blood-gas measurements (Table 1). The qualitative categories represent linguistic terms often used in daily clinical practice.

These seven regions of interest are not equal sized. The value range of an interval is smaller the nearer the target range. This is an important feature representing dynamics related to the different degrees of parameters' abnormalities. It is extensively used in the schemata for trend-curve fitting discussed in the next section.

The schemata for data-point transformation are defined for all kinds of blood-gas measurements. They further depend on the clinical context, resulting in different numerical ranges. Figure 1 shows the schemata for transformation of $P_{tc}CO_2$. On the left-hand side the scheme for transformation during intermittent positive pressure ventilation (IPPV) is shown, on the right-hand side the scheme for transformation during intermittent mandatory ventilation (IMV). Depending on the clinical context (IPPV or IMV) the same numerical reading results in different qualitative statements.

The data-point transformation schemata transform each (valid) numerical value into one of seven qualitative categories. This temporal abstraction pro-

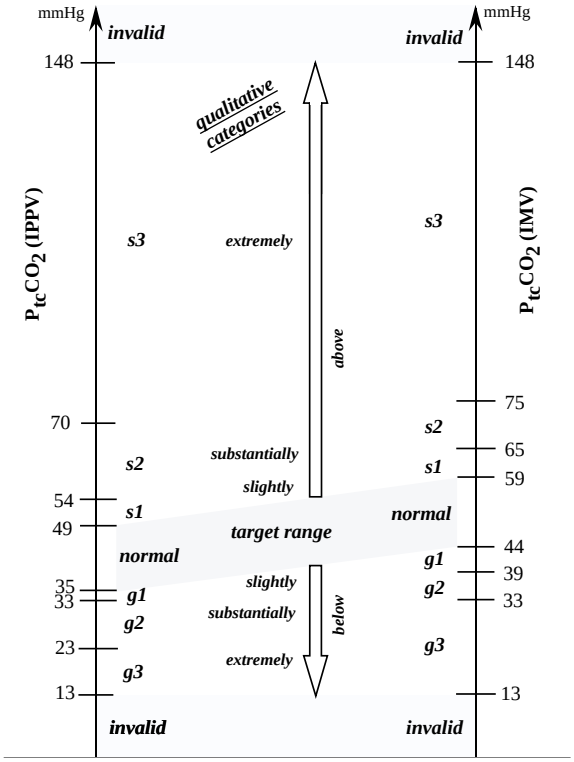

Fig. 1. Schemata for data-point transformation of $P_{tc}CO_2$ during context intermittent positive pressure ventilation (IPPV, left) and intermittent mandatory ventilation (IMV,right). The qualitative data point categories are given in the middle column. For example, a $P_{tc}CO_2$ value of 60 will be transformed to "substantially above target range (s2)" during IPPV and to "slightly above target range (s1)" during IMV.

cess is further enhanced by three more options:

(i) Smoothing of data oscillating near thresholds. If a data value moves out of a qualitative region a little bit for just a few seconds the qualitative category does not change. This avoids rapid changes of qualitative categories in cases the numerical value oscillates around a threshold.

(ii) Smoothing of data-point transformation schemata. This supports a graceful change when changing the clinical context (e.g., changing the mode of ventilation from IPPV to IMV). Figure 1 gives such an example: during several hours the IPPV scheme on the left is changed gradually to the IMV scheme on the right. Without such a smoothing, e.g., a $P_{tc}CO_2$ value of 55 would change within a second from "s2" to "normal" causing severe changes in the treatment. This has to be avoided.

(iii) Context-sensitive adjustment of qualitative values. This allows to adjust qualitative values during life-threatening situations to be able to tolerate higher values as better ones under specific circumstances. Specific rules are used to modify a qualitative category in case of life-threatening situations (e.g., very high peak inspiratory pressure).

Table 2
The four kinds of trends, the duration used to compute the trends, and the criteria for determining the validity of the trends.

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

These enhancements are discussed in detail in [17].

4.2 Expected Qualitative Trend Descriptions

Similar to the transformation of numerical data points to qualitative values, interval data are transformed to qualitative descriptions resulting in a verbal categorization of the change of a parameter over time. Analogous to the data point transformation scheme we build an unified scheme for these qualitative trend descriptions. The transformation of interval data to qualitative trend descriptions is an abstraction process, which needs to adapt to the dynamics of the continuously-assessed parameters. This data-abstraction process builds dynamically-derived qualitative trend categories, which overcome the limitations of predefined static thresholds.

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 the time-ordered data they use to calculate the trend. Further, they differ in the validity criteria, which have to be fulfilled to be able to determine a valid trend. In monitoring more recent data are more important compared to older measurements. Due to this precondition 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 can easily be adapted to a specific clinical situation based on the frequency at which data values arrive. Table 2 summarizes the trends and their criteria. For each kind of trend the actual growth rate and the derived qualitative trend category is determined.

The trend-abstraction process is based on *expected qualitative trend descrip-*

Table 3
The ten qualitative trend categories derived from the actual growth rate of a parameter. Categories A2 and B2 represent the expected normal change of a parameter if it is above or below the target range, respectively. The categories are based on the expectation that a parameter should return to its target range within a time period which one expects from physiology.

Region	Code	Trend Category
upper	C	dangerous increase
	ZA	zero change
	A3	decrease too slow
	A2	normal decrease
	A1	decrease too fast
lower	B1	increase too fast
	B2	normal increase
	B3	increase too slow
	ZB	zero change
	D	dangerous decrease

tions. These are qualitative statements, which express physicians' expectations for how a blood-gas value has to change over time to reach the target range in a physiologically proper way. For example, "the parameter $P_{tc}CO_2$ is moving one qualitative step towards the target range within 20 to 30 minutes" is such a statement related to the expectation of "normal decrease". Qualitative steps are defined in terms of the qualitative data-point categories of section 4.1, which are sized differently. Applying the expected qualitative trend descriptions to these data-point categories, we get a qualitative notion of "normal decrease" for the upper region and "normal increase" for the lower region. Both are defined by a specific area of growth. The assumed exponential functions, which delimit these areas, are determined through stepwise linearization and a dynamic comparison algorithm. This reduces complexity considerably. The comparison algorithm utilizes a *trend curve fitting scheme* to transform a growth rate into one of ten qualitative trend categories. The categories are divided by the target range into an upper and a lower region (Table 3).

Details about the trend-curve-fitting scheme are in [9]. The trend categories define a partial ordering for both the upper and the lower region. This partial ordering allows to use the qualitative categories for data validation (see section 5.3).

Figure 2 shows the principle method taking the actual development of $P_{tc}CO_2$ during IMV as an example. What we would like to achieve are trends, which

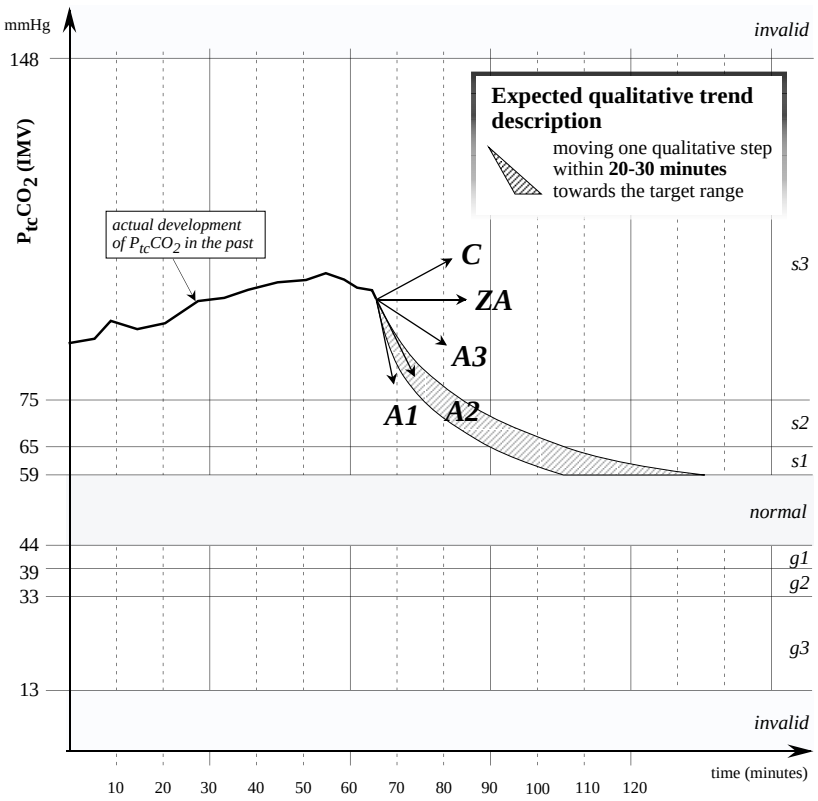

Fig. 2. Principle method to qualify the actual development of a continuously-assessed parameter. The striped area A2 shows the expected normal decrease towards the target region. The development of a parameter in the past (in the example $P_{tc}CO_2$ during IMV) is abstracted to qualitative trend categories (written in bold, capital letters).

go into the A2 direction, namely a normal decrease towards the target region. Evaluating the actual development of $P_{tc}CO_2$ in the example we observe a dangerously increasing long-term trend (last 3 hours, category C), a nearly zero change of the medium-term trend (last 30 minutes, category ZA), a decrease too slow of the short-term trend (last 10 minutes, category A3), and a normal decrease of the very-short-term trend (last minute, category A2).

A brief explanation about how the very-short-term trend is determined to be A2 should clarify the use of the trend curve fitting scheme. We explain it for values in the upper region (above the target range). The current value of $P_{tc}CO_2$ is 97 mmHg, the qualitative value of $P_{tc}CO_2$ is $s3$. Let k_a be the actual growth rate of $P_{tc}CO_2$. k_a is calculated applying a simple linear regression model on the valid $P_{tc}CO_2$ data values of the last minute. The *expected qualitative trend description* for $P_{tc}CO_2$ specifies the change of "one qualitative step in 20 to 30 minutes". One qualitative step is the change from the current value 97 in qualitative region $s3$ to the corresponding value in qualitative region $s2$, i.e., 68 mmHg. Let k_1 be the growth rate resulting from a decrease of $P_{tc}CO_2$ from 97 to 68 in exact 20 minutes. Let k_2 be the equivalent

growth rate for taking 30 minutes to reach 68 mmHg. If k_a is in the interval $[k_1, k_2]$ the qualitative very-short-term trend belongs to category A2 ("normal decrease"). For $k_a < k_1$ category is A1, for $k_a \in (k_2, -\varepsilon)$ category is A3, for $k_a \in [-\varepsilon, \varepsilon]$ category is ZA, and for $k_a > \varepsilon$ category is C. ε is used to define the "region of zero change" ZA. In our example the k_a is in the $[k_1, k_2]$ interval. As a result the very-short-term trend category is A2.

The comparison algorithm behaves similar for data values below the target range. For the short-term, medium-term, or long-term trend the regression analysis which determines k_a uses valid data values of the last 10, 30, or 180 minutes, respectively. The two thresholds k_1 and k_2 vary for each data value of $P_{tc}CO_2$, basically due to the increasing size of the qualitative regions $s1$, $s2$, and $s3$. In summary this results in a stepwise linear approximation of an assumed exponential improvement of the parameter.

The data-selection procedure of VIE-VENT returns a data value for each of the continuously-assessed parameters once per second. Data abstraction gives their derived context-sensitive, qualitative data-point categories, and the growth rates of the different trends and their derived qualitative trend categories. Data validation described in the next section use these continuously-assessed data together with the discontinuously-assessed parameters. It tries to produce a consistent overall view, which of the parameters are valid and which are not.

5 Data Validation and Repair Methods

We distinguished three categories of data validation and repair based on their underlying temporal ontologies: time-point-, time-interval-, trend-based validation and repair. Further, a time-independent validation method is used to rate the reliability of specific parameters. Table 4 gives an overview of all methods applied. The methods are grouped by the underlying ontology and qualified by the kind of data used (quantitative or qualitative) and the action performed (validation or repair).

The basic functionality of the methods and their use is as follows:

- *Range checking* determines if a quantitative value is within an acceptable range. Both, data points and trends can be checked.
- *Causal dependencies* check for the existence of relationships between parameters. Parameters can be quantitative or qualitative.
- *Functional dependencies* extend the previous check by establishing explicit functions between parameters. Functional dependencies are used to compare quantitative and/or qualitative data-point values. Further, they com-

Table 4
Data validation and repair methods

Ontology	Method	quant.	qual.	validat.	repair
Time-point	Range checking	x		x	
	Causal dependencies	x	x	x	
	Functional dependencies	x	x	x	x
Time-interval	Temporal validity	x		x	x
	Stability check	x		x	
	Cross-validation	x		x	
	Dynamic calibration	x			x
Trend	Range checking	x		x	
	Højstrup modified	x		x	
	Functional dependencies		x	x	
	Trend assessment		x	x	
	Predicting values	x	x		x
Time-independent	Priority lists		x	x	x

pare qualitative trends of different but related parameters. In addition, the functions are used to repair unknown or invalid values.

- *Temporal validity* determines the interval a parameter is valid or invalid.
- *Stability check* enforces a certain time period a parameter has to be stable. This interval of stability is required for an invalid parameter to become valid.
- *Cross-validation* checks for consistency of different parameters during a given time interval.
- *Dynamic calibration* repairs invalid values during a time interval. It applies a calibration function to a continuously-assessed parameter utilizing a valid (discontinuously-assessed) parameter. The assumption is that the invalid continuously-assessed parameter has a valid trend.
- *Højstrup method* checks for growth rates which are too steep. Depending on the size of the growth rate the invalidity is signalled immediately or after some time period.
- *Trend assessment* examines successive qualitative trend values. If they differ too much the corresponding parameter becomes invalid.
- *Predicting values* is a method for determining values of unknown or invalid parameters. It uses trends to set values for these parameters.

In principle, the order of listing the methods expresses the order of their application. The order of application is of specific importance both for the

data abstraction process and the data validation. Details about ordering are discussed in section 5.5. The methods are presented in detail in the following sections 5.1 to 5.4.

5.1 Time-Point-Based Validation and Repair

The time-point-based category uses for the reasoning process the value of a variable at a particular time point. This concept can handle all kinds 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 applied range checking as well as causal and functional dependencies to detect faulty values. We extended the concept of functional and causal dependencies to deal with qualitative functional dependencies and with inaccurate measurements caused by measuring faults. Invalid values are repaired by applying functional dependencies or using a simplified model, which is able to cope with missing values.

Range checking

Range checking is basically simple but has shown very powerful to detect disconnections and missing measurements. Figure 1 gives an example by marking the invalid regions for the transcutaneous measurement $P_{tc}CO_2$. Most modern equipment is able to perform range checks by itself, but most often the data available at serial or analog lines do not include the information whether the data point is within the chosen range or not. Further, range checks have to be sensitive to the explainable errors caused by A/D conversion and the precision of the instruments. Most often we have received a S_aO_2 reading of 100.1%, which would be classified as invalid without inclusion of an explained error.

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., $tidal\,volume \leq 5ml/kg$). A causal dependency can be bidirectional—as shown in the example above—or unidirectional. In the bidirectional case we are able to conclude that some of the parameters are wrong. The unidirectional case allows to invalidate a specific parameter. For example, S_aO_2 is invalidated if we can't 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 $beats/min$):

$$valid(PULS) = false \rightarrow valid(S_aO_2) = false \tag{1}$$

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

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.

Functional dependencies

Functional dependencies are useful both for numerical and qualitative parameters. Applying a functional dependency not only provides a mean for validating the parameters of the function, but gives a way to repair 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. E.g.,

$$f = \frac{60}{t_i + t_e} \quad (3)$$

relates frequency f with inspiration time t_i and expiration time t_e. Most important, rounding errors and errors resulting from A/D conversion (explained error) does not allow to use the exact equation (3), but forces to compare the real difference between the left and right side of the equation with the maximum allowed difference due to the explained error of the parameters.

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 4.1 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 category, 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 *s2* 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 priority list discussed in section 5.4 and the dynamic reliability score computed by each of the various validation methods. Lateron in the validation process we will either invalidate one of the two measurements or repair it using dynamic calibration.

Comparing transcutaneous and invasive blood-gas measurements involves another need for a special management of time-stamped data: time-synchronization of the measurements. Taking an invasive blood-gas sample at timestamp t_x with the results available after some minutes, say at timestamp t_{x+n}, we have

to remember the $P_{tc}CO_2(t_x)$ and compare it with the $P_cCO_2(t_{x+n})$. This can result in the necessity of revising past decisions. We neglect this due to the impossibility of changing recommendations already given, but we use it correctly for time-interval-based cross-validation and repair discussed in the next section.

5.2 Time-Interval-Based Validation and Repair

The time-interval-based category deals with the values of different variables within a time interval. We used three methods: (1) temporal validity of measurements, (2) allowed changes of values of a single variable depending whether a therapeutic action has taken place, (3) cross-validating data from different sources (e.g., continuously and discontinuously-observed data). We applied a dynamic calibration of values acquired by different sources to repair invalid values.

Temporal validity

Temporal validity sets the time interval a parameter is valid. For discontinuously-assessed data there are two possibilities for setting the valid time interval:

- The user of VIE-VENT can specify the duration of validity when entering a particular discontinuous data value. E.g., "P_aO_2 should be valid for the next 30 minutes".
- 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:

- the time interval of the parameter's validity has elapsed,
- a new value of the parameter is available, or
- an external event enforces to manually set the parameter invalid.

The reliability score of a discontinuous parameter gets smaller over time. The temporal validity interval determines how long the time-interval-based repair method *dynamic calibration* can be active.

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).

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 method defines allowed changes of 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:

- 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.,

$$\forall i, i = 1, \ldots, n : \mid S_aO_2(t) - S_aO_2(t+i) \mid \leq \varepsilon \tag{4}$$

  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.
- 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.

Cross-validation

Cross-validation of data from different sources is the time-interval-based utilization of qualitative functional dependencies described in section 5.1. 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.

In 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.

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 is applied during the *temporal validity interval* of X. The resulting repaired value of Y receives a decreasing reliability score over time.

Dynamic calibration is motivated by an initial study which shows a strict correlation between tanscutaneous and arterial measurements in children of age four months to 14 years [18]. In a more recent independent study we analysed 442 cases with corresponding carbon dioxide measurements measurements at a neonatal ICU using regression analysis. We found a function which correlates the arterial (P_aCO_2) and transcutaneous ($P_{tc}CO_2$) measurements with an acceptable correlation coefficient r:

$$P_{tc}CO_2^{corr} = 2.226 + 1.039 P_aCO_2, r = 0.839 \tag{5}$$

We were not able to find an acceptable correlation function for invasive and transcutaneous oxygen measurements. One possible explanation for this missing correlation is the bad circulation of neonates.

Dynamic calibration uses equation 5 to repair transcutaneous carbon dioxide measurements in case cross-validation signals a significant difference between P_aCO_2 and $P_{tc}CO_2$. If dynamic calibration is initiated at time point t_x and PCO_2^{meas} are the measured values, we calculate calibrated $P_{tc}CO_2^{cal}$ for each time point $t_y = t_{x+m}$:

$$\begin{aligned} P_{tc}CO_2^{cal}(t_y) = {} & P_{tc}CO_2^{meas}(t_y) + \\ & P_{tc}CO_2^{corr}(t_x) - P_{tc}CO_2^{meas}(t_x) \end{aligned} \tag{6}$$

The calibration is done for each m in the temporal validity interval of $P_aCO_2(t_x)$.

5.3 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 categories (descriptions) presented in section 4.2. We applied (1) range checks on the growth rate, (2) an evaluation procedure, which inspects the temporal behavior of measurements (Højstrup method modified), (3) trend-based functional dependencies

of different dependent variables, and (4) an assessment procedure of the development of a variable. Predicting values is a repair method with deals with missing values.

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 react thus very fast.

Højstrup method modified

The Højstrup method modified recognizes growth rates, which are unacceptable after a certain amount of time. It recognizes unplausible 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 unplausible 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 [19]. We have modified the algorithm to the needs of real-time monitoring in ICUs.

Starting with a sequence of data points $x_0, x_1, \ldots, x_{i-1}$, the mean value m of this sequence, and the correlation K of two neighboring points it predicts the next data value v_i by

$$v_i = x_{i-1}K + (1 - K)m \tag{7}$$

Getting the new value x_i the mean m and the correlation K will be updated. Based on the assumption that the difference between the predicted v_i and the actual measured value x_i follows a Gaussian distribution a threshold for this difference is defined. E.g., the error threshold E can be fixed to a value that the probability is less than 0.01 that a correct value exceeds the threshold.

The algorithm has been modified to the requirements of analyzing blood gas values: the correlation function K is replaced by a measurement for the deviation of the last two points from the mean. We further can 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 works as follows:

(i) Using the last measurement x_{i-1}, the last mean m_{i-1}, and the last devi-

ation s_{i-1} predict the next value v_i:

$$v_i = x_{i-1} e^{\frac{-|s_{i-1}|}{R}} + m_{i-1}(1 - e^{\frac{-|s_{i-1}|}{R}}) \tag{8}$$

(ii) Get the new data value x_i

(iii) Update weighted mean and deviation measure:

$$m_i = m_{i-1}(1 - \frac{1}{M}) + \frac{x_i}{M} \tag{9}$$

$$s_i = s_{i-1}(1 - \frac{1}{M}) + \frac{(x_i - m_i)(x_{i-1} - m_{i-1})}{M} \tag{10}$$

(iv) Decide whether x_i is valid:

$$| v_i - x_i | > E \rightarrow valid(x_i) = false \tag{11}$$

(v) Continue with next i.

The important parameters of this algorithm are M and R. They influence on the one hand the calculation of the predicted value v_i and the update of the mean m_i and the deviation s_i, and on the other hand the classification of unplausible values. The weight M determines how strong old x values influence the calculation. For eaxmple, $M = 2$ will update $m_i = (m_{i-1} + x_i)/2$. On the extreme, $M = 1$ will ignore m_{i-1} and set $m_i = x_i$, whereas $M \rightarrow \infty$ will keep a constant m_i. R determines the shape of the exponential curve used to determine the predicted values v_i. By these means they specify which growth rate is "too steep". Therefore, the fine tuning of the determining parameters M, R (and E) is the critical part of the algorithm.

A systematic analysis was performed modeling the algorithm in the form of a RC-low-pass filter. Equations 9 and 10 have been expressed as differential equations and the limits to their solutions have been analyzed. Using this analysis it is possible to derive values for M, R, and E from three plausible parameters: the sampling rate T, the steepest growth which is valid $T_{a_{min}}$, and the rigidity Rig. Let T_a be a measurement for the growth rate by counting the number of data points until the data value has changed by one unit. For a fixed T_a the difference between the measured values and the calculated means converges towards a boundary value. The predicted values follow the slope of measurements with an error that depends on T_a. The error converges towards a boundary value $E_b(T_a)$. The smaller T_a is the greater is $E_b(T_a)$. For a fixed E we can approximate the time $T_g(T_a)$, which is needed until $E_b(T_a)$ is greater than E. There is a logarithmic relation between growth rate T_a and the time needed to signal an error $T_g(T_a)$. $T_{a_{min}}$ is that T_a where it takes infinite time to signal an error. For all $T_a > T_{a_{min}}$ we will never receive an error. Further, there exists a $T_{a_{imm}}$, where $E(T_{a_{imm}})$ exceeds E within the time until the next data point is taken. Thus, the error is signaled immediately. For T_a in between

$T_{a_{imm}}$ and $T_{a_{min}}$ the rigidity Rig determines $T_g(T_a)$. The higher the rigidity, the faster an error will be reported. Details are to be found in [20].

Several experiments have been done to get the desired behavior for each of the continuously-assessed parameters. For example, a sampling rate of 60 data points per minute ($T = 60$), the acceptance of a change of 6 units per minute ($T_{a_{min}} = 10$), and a low rigidity ($Rig = 0.2$) result in $M = 1.3, R = 0.39, E = 7.6$. These values have shown useful for validating S_aO_2. They produce a validation behavior similar to expert neonatologists. Values showing stability with slight variations are accepted. Values above the 6 units per minute change rate limit are rejected immediately. Values which define changes rates below this limit are marked invalid if they persist for some time.

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.

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 4.2 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 ambiguous.

A second usage of trend-based functional dependencies is the validation 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 FiO_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 utilization of trend-based functional dependencies after a therapeutic action has to consider two complications: (1) if the trend of the monitor parameter does not change, the system has to decide whether the parameter is invalid or the therapy is not effective. Therefore, the method is useful for both data validation and therapy assessment; (2) during the delay time needed for an effect of a therapeutic action to become visible other therapeutic actions can

cause adverse effects to the parameter in question. For example, a decrease of the peak inspiratory pressure PIP can adverse the increase of O_2 caused by the FiO_2 increase. We have to take into account all known influences during the time interval, which is composed of the delay time and the time needed to compute the trend.

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 will assume $P_{tc}O_2$ giving invalid readings due to some other causes, like bad circulation.

Trend assessment

The assessment procedure of the development of a parameter 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.

Figure 3 gives an in-depth example of the assessment procedure for the parameter $P_{tc}CO_2$. It plots 33 minutes of $P_{tc}CO_2$ (given in mmHg) plus the qualitative data-point values and derived qualitative trend categories in the corresponding columns of the tables. Inspecting the short-term trend we detect a change from $A3$ to $A1$ at 20:57. As these two are not neighboring categories, an error is signaled and the $P_{tc}CO_2$ value is invalidated. The figure shows further that the very-short-term trend reacts too fast. The rapid changes between qualitative categories (e.g., at 20:45) are not useful for detecting invalid or suspect measurements.

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.

Predicting values

During a monitoring process the position of a sensor has to be changed frequently and regularly. Therefore, the measurements are often missing. The

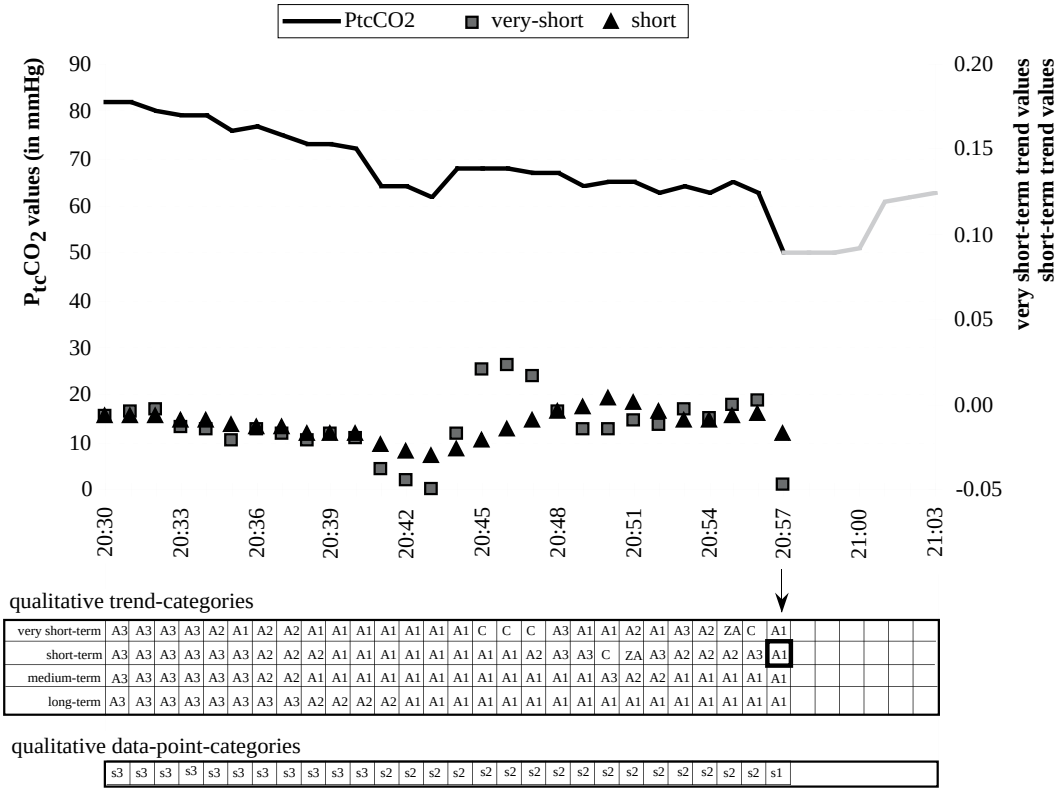

Fig. 3. Example assessing the development of the parameter $P_{tc}CO_2$. The x axis represents the time in minutes. The left y axis gives the $P_{tc}CO_2$ value in mmHg, the right y axis measures the very-short-term and short-term trends. The tables below the plot show the corresponding qualitative trend categories and data-point categories, respectively. The faulty measurement is detected at 20:57 using the short-term trend category.

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. Second, applying the growth rate of the short-term trend to predict a "correct" value. A precondition is the stability of the trend. The stability is assessed applying the qualitative trend-categories. If the medium-term and short-term qualitative trend-categories are identical, the precondition of intrinsic development of the measurements becomes true. The trend-based prediction of a value is a more accurate action, because it takes the history of the values into account. But the criteria of validity to calculate a trend have to be fulfilled to predict a value.

Predicting 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 are not able to get valid measurements

over a longer period of time, VIE-VENT falls back to a simplified reasoning process.

The simplified reasoning process uses only a few parameters. VIE-VENT uses a simplified system model of neonatal respiration during the initial phase when the only reliable continuous measurement is S_aO_2. There are restricted reactions to decrease oxygenation depending on the degree of abnormality of the S_aO_2 and the actual tidal volume VT. The tidal volume is estimated here by the extent of the chest wall expansion.

5.4 Time-Independent Validation

The last category is based on time-independent priority lists of variables. Priority lists of the measurements are an indicator of the reliability of measurements. The data-validation process identifies a less reliable parameter from a set of conflicting parameters. The result is a reliability ranking. 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. 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.

Examples of priority lists of VIE-VENT are: arterial blood gases are more reliable than venous blood gases; invasive blood gases are more reliable than transcutaneous blood gases and they are more reliable than S_aO_2; and S_aO_2 is more reliable than $P_{tc}O_2$.

5.5 Interaction of the Methods

The sequence of presentation of data-validation methods above defines the principle sequence these methods are applied. The reasoning methods based on time points and time intervals represent a preprocessing for the reasoning based on trends. They primarily perform static data validation which delivers the necessary preconditions to proceed with the trend-based validation.

Temporal data abstraction is a prerequisite for all data-validation methods operating on qualitative data. Thus, a first step is the abstraction of time-point-based data. Next we apply data-validation methods based on time points and time intervals. If repair is needed, we will further use the time-point- and time-interval-based repair methods. In the next step we calculate trends for all parameters found valid by the previous steps, both by linear regression methods and by determining the qualitative trend categories. Next, we apply

trend-based data validation and repair. The priority list of parameters and the dynamic reliability score of the parameters is consulted if there is a necessity to decide towards a more reliable parameter in case of ambiguity.

Trend-based validation can result in the conclusion that a data value of the last time point is implausible and has to be invalidated (in some severe cases even older values have to be invalidated). In such a case previous validation methods have to be reapplied.

The sequence of application of methods is further complicated by the causal and functional dependencies of the parameters. The example of the causal relation between heart rate HR from ECG, pulse $PULS$ from pulsoximetry, and S_aO_2 given in equations (1) and (2) demonstrates the complexity in scheduling the validation process. First we have to use all known methods to validate HR and $PULS$. This is resolved by explicitly representing dependencies of parameters and methods.

Given such strong interaction of the methods presented the data-abstraction and the data-validation processes have been implemented in a modular form. The modules are activated whenever applicable resulting in a multi-step procedure.

6 Discussion

VIE-VENT integrates data validation with data abstraction. This two processes are strongly intertwined due to the need of abstracted data values and trends for high-level data validation, and due to the need of valid data for the abstraction process.

Looking at data abstraction, several significant and encouraging approaches have been developed in the past years: RÉSUMÉ [15,16] supports temporal abstraction of time-stamped data. It performs context-dependent temporal abstraction and temporal reasoning over intervals. However, the approach suffers from several limitations: RÉSUMÉ covers only limited domain dynamics. It does not include different classifiers for different degrees of parameters' abnormalities. It requires predefined domain knowledge to perform temporal interpolation (e.g., gap functions), which is not available in some domains. It is mainly designed to cope with low-frequency observations which cannot easily be adapted for high-frequency data due to their different properties. Finally, different contexts have to be defined in advance and are not automatically deduced from the input parameters.

TrenDX [21] detects clinically significant trends in series of time-ordered data.

It uses trend templates to define disorders as typical patterns of relevant parameters. These patterns consist of a partially ordered set of temporal intervals with uncertain endpoints. The drawbacks of this method lie in the predefinition of the expected normal behavior of parameters during the whole observation process and the usage of absolute value thresholds matching a trend template. The absolute thresholds do not take into account the different degrees of parameters' abnormalities. In many domains it is impossible to define such static trajectories of the observed parameters in advance. Depending on the degrees of parameters' abnormalities and on the various contexts, different normal behaviors are expected. These normal expectations vary according to the patient's status in the past. Therefore these thresholds have to be derived dynamically during the observation period. For example, the decreasing of transcutaneous partial pressure of carbon dioxide ($P_{tc}CO_2$) from 97 mmHg to 91 mmHg during the last 25 minutes would be assessed as "decrease too slow" because the patient's respiratory status was extremely above the target range in the past. However, the same amount of change (6 units) from 64 mmHg to 58 mmHg would be assessed as "normal decrease" during a period where the patient's respiratory status was slightly above the target range.

The main focus of temporal data abstraction is to provide short, informative, and context-dependent summaries of time-oriented data. This goal can be achieved by the elimination of unimportant details (e.g., the Temporal Control System TCS [22]), by creation of synthetic views of the patient's clinical history (e.g., M-HTP [23]), or by assessing the evolution of the patient's status [24]. A comprehensive review of temporal-reasoning approaches is given in [16]. All these systems use data abstraction to support a higher level of clinical reasoning. This improves monitoring and therapy planning.

VIE-VENT extends the approach of data abstraction to utilize the derived temporal patterns for data validation as well as for monitoring and therapy planning. Its data-abstraction mechanism is well suited for domains with high-frequency data acquisition. Its abstraction process is context-sensitive and expectation-guided. This allows dynamic orientation to the clinical context. The behavior of parameters over time is qualified by an expectation-guided principle. It is based on the experience of domain experts about the expected normal development of parameters of critically ill neonates. The outcome of VIE-VENT's data-abstraction process are unified qualitative values for data points and trends, easy to comprehend for experts and easy to use for data interpretation and therapy planning.

VIE-VENT receives data from monitors once per second. Its efficient temporal representation and its understanding of the temporal characteristics of the monitored parameters enable VIE-VENT to operate in real-time. This efficiency in temporal representation is to be seen as one of the prerequisites for completing each cycle of data validation, data abstraction, and data interpre-

tation in real-time [4].

Looking at data validation one direction is the effort to find complete data sets containing reliable data [5]. This is motivated basically by data collection procedures, which operate on remote sites with data input from clinical personnel. The range checks and referential integrity conditions are useful specifically for discontinuously-assessed parameters. Continuously-assessed parameters—received on-line—require methods which are much more enhanced. For example, it is necessary to include the explained error (of a sensor and the A/D converter) in range checks and functional dependencies to be able to accept a reading of $100.1\% FiO_2$ as valid.

In high-frequency domains artifact-recognition methods include statistical signal processing techniques and neural networks. Statistical signal processing, like Kalman filtering, is computationally expensive [25]. 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 [26]. The second main area of artifact recognition is intelligent alarming. It has shown useful in anesthesia monitoring [27] and post-operative care [28]. The combination of range checks and validation and invalidation rules provide good results in eliminating false alarms [29]. For this purpose simple rules like the "small change rule" *(if the prior value was valid and the new value showed a 7% or smaller change, then the new value is valid)* are most useful. In effect, VIE-VENT's expectation-guided trend templates cover such concepts in an unified fashion, which is based on the knowledge of experts.

VIE-VENT's data validation methods include limited repair facilities. The repair methods find plausible values for continuously-assessed parameters which are unknown or invalid at moment. Recent techniques of time-series analysis and forecasting, like probabilistic belief networks [30] and dynamic network models [31], tries to find a proper probabilistic model under the assumption of some regularities of the underlying dynamical processes. Such computational expensive methods are not applicable in the absence of a structure-function model, as in the domain of neonatal respiration. In contrast, VIE-VENT concentrates on getting a consistent view by combining all information available. It tries to find the parameters most reliable and to use these parameters and known trends to derive values for the parameters currently unknown or invalid. Dynamic calibration is a useful method to repair unreliable continuously-assessed measurements by using reliable discontinuously-assessed measurements under the assumption of a reliable trend of the continuously-assessed data. VIE-VENT's dynamic scoring scheme is able to judge the reliability of each parameter at each time point. It combines the static reliability of a parameter stored in priority lists with dynamic reliability scores computed by each of the various validation methods.

7 Evaluation

To evaluate VIE-VENT's data validation procedures we created an evaluation scenario. This scenario consists of two steps: a visual inspection of the results of the data validation process, and a formal evaluation of the validation results. The evaluation of validation procedures for continuously-assessed data is hard due to lack of a gold standard. Therefore we have to rely on the judgement of experts, experienced neonatologists in our case.

Step one of the evaluation procedure is based on the visual inspection of the results of the data validation procedure by experts. The judgements of the experts are based on two outputs of VIE-VENT: (1) the visualization of the data points and data sequences found invalid and (2) VIE-VENT's overall therapeutic recommendations given every 10 seconds. Therapeutic recommendations are important because the data validation process has to invalidate short variations in the continuously-assessed data. Such short-term variations would result in incorrect therapeutic recommendations. For validation we used 640786 seconds (approx. 178 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. We have taken the data from the first six patients (approx. 115 hours) to tune the validation parameters, specifically to find suitable parameters for the stability check and the Højstrup method. All other validation parameters were derived from the knowledge of expert neonatologists. The results of the data validation are visualized by marking invalid data with rectangular markers in the bottom of each curve. The markers appear as stripes for continuous sequences of invalid data points. Two expert neonatologists examined the results. The validation parameters have been adjusted until accepted by the experts. The remaining data (approx. 62 hours of recording) have been used to verify the correctness of VIE-VENT using the same visual inspection by our experts. Figure 4 shows an example plot of a 60-minutes sequence of on-line recording including the invalidation markers. The figure gives S_aO_2 (in %), $P_{tc}O_2$ (mmHg), $P_{tc}CO_2$ (mmHg), heart rate from ECG and pulse from pulsoximetry. The invalidation markers of the upper three plots show the data points classified invalid by VIE-VENT.

Data validation in VIE-VENT is tuned towards the overall goal to avoid wrong therapeutic recommendations. As a consequence data validation tries to mark all measurements as invalid which are *unusable* signals. The principle goal is marking signals as *unusable* if they will result in short-term variations of therapeutic recommendations. An example is the stability check. It marks a parameter valid only if it is stable for at least two minutes. This extends the usual methods of identifying artifacts.

The second step in our evaluation scenario is a formal evaluation. We compared

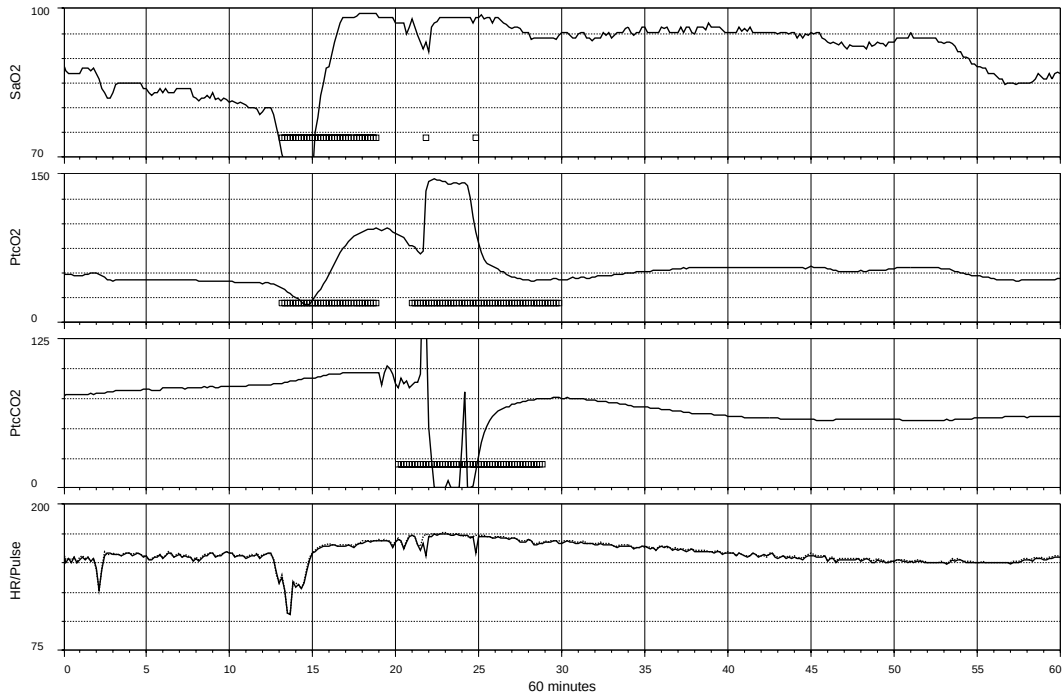

Fig. 4. Sample monitoring results showing S_aO_2 (in %), $P_{tc}O_2$ (mmHg), $P_{tc}CO_2$ (mmHg), heart rate from ECG and pulse from pulsoximetry (overlayed in the bottom plot). Stripes and rectangles near the x axis of each plot mark the data points classified invalid by VIE-VENT.

the data points found invalid by an expert neonatologist with the invalidation markers produced by VIE-VENT. In the absence of a gold standard for evaluating sequences of continuously-assessed recordings of $P_{tc}O_2$, $P_{tc}CO_2$ and S_aO_2 we use the expert's decision as the standard. For this evaluation study we have taken sequences of continuously-assessed data which show some variation. If there is no variation in the data values it is obvious that both the expert and VIE-VENT will judge the data being valid. In this case we will receive always true negatives in comparing the expert and VIE-VENT. We have 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 5 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 the expert is able to see the future development of a parameter from the plot. In contrast, VIE-VENT operates in real-time. It

Table 5
Evaluation of VIE-VENT's data validation procedures by using the judgements of an expert neonatologist as the standard.

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

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 required for real-time operation.

VIE-VENT is designed for real-time operation at neonatal ICUs. Given such a complex clinical environment a set of methods is required operating on both numerical input and qualitative values, which are either manual input values or values derived from data-abstraction procedures. The effectiveness of data validation in VIE-VENT results from the combination of a diversity of methods. The enhanced possibilities to cross-validate parameters and to check functional dependencies of both data points and trends provide the robust basis to eliminate measurements which would result in wrong therapeutic recommendations of VIE-VENT.

8 Conclusion

We demonstrate methods for automated data validation and repair based on different temporal ontologies (time points, time intervals, and trends). They take into account the various types of available data occurring at various frequencies. They combine and integrate a set of methods for data validation in a real-time high-frequency environment. It is important to use all available information for data validation, to cross-validate continuously-observed and discontinuously-observed data, and to cross-validate data from different sources. Of essential importance is the reliability ranking of data values to reach meaningful conclusions in conflicting situations. Such reliability can result from a priori definitions, from experience, or from dynamic evaluation of the current data set.

Our approach benefits from dynamically-derived qualitative data-point and trend categories which result in unified qualitative descriptions of parameters and overcome the limitations of comparison with predefined static thresholds.

Applying our validation methods to the observed on-line and off-line data sets resulted in automatic elimination of invalid measurements. Using these classified measurements improved the monitoring and the therapy-planning process

significantly: (1) false positive alarms were reduced, (2) errors of data interpretation were minimized, (3) abrupt changes of therapeutic recommendations were eliminated promoting a stable and graceful weaning process.

Acknowledgement

The project was supported partially by the "Jubiläumsfonds der Oesterreichischen Nationalbank", Vienna, Austria, project number 4666. Current research is supported by "Erwin Schrödinger Auslandstipendium, Fonds zur Förderung der wissenschaftlichen Forschung", J01042-MAT. We greatly appreciate the support given to the Austrian Research Institute of Artificial Intelligence (OFAI) by the Austrian Federal Ministry of Science, Transport, and the Arts, Vienna.

References

[1] R.M. Gardner, B.J. West, T.A. Pryor, K.G. Larsen, H.R. Warner, T.P. Clemmer and J.F. Orme, Computer-Based ICU Data Acquisition as an Aid to Clinical Decision-Making, *Crit. Care Med.* **10** (1982) 823–830.

[2] S.T. Lawless, Crying Wolf: False Alarms in a Pediatric Intensive Care Unit, *Crit. Care Med.* **22** (1994) 981–985.

[3] S. Uckun, Intelligent Systems in Patient Monitoring and Therapy Management, *Int. J. Clin. Mon. Comp.* **11** (1994) 241–253.

[4] D.J. Musliner, J.A. Hendler, A.K. Agrawala, E.H. Durfee, J.K. Strosnider and C.J. Paul, The Challenges of Real-Time AI, *IEEE Comp.* **28** (1995) 58–66.

[5] D. Carlson, J. Wallace, T.D. East and A.H. Morris, Verification and Validation Algorithms for Data Used in Critical Care Decision Support Systems, in R.M. Gardner, ed., *Proceedings SCAMC'95* (New Orleans, Louisiana, 1995).

[6] R.M. Gardner, W.L. Hawley, T.D. East, T.A. Oniki and H.-F.W. Young, Real Time Data Acquisition: Experience with the Medical Information Bus (MIB), in: P.D. Clayton, ed., *Proceedings of the Fifteenth Annual Symposium on Computer Applications in Medical Care, SCAMC-91* (McGraw-Hill, New York, 1992).

[7] S. Miksch, W. Horn, C. Popow and F. Paky, VIE-VENT: Knowledge-Based Monitoring and Therapy Planning of the Artificial Ventilation of Newborn Infants, in: S. Andreassen et al., eds., *Artificial Intelligence in Medicine, Proceedings AIME-93* (IOS, Amsterdam, 1993) 218–229.

[8] S. Miksch, W. Horn, C. Popow and F. Paky, Context-Sensitive Data Validation and Data Abstraction for Knowledge-Based Monitoring, in: A.G. Cohn, ed., *Proceedings of the 11th European Conference on Artificial Intelligence, ECAI94* (Wiley, Chichester, 1994) 48–52.

[9] S. Miksch, W. Horn, C. Popow and F. Paky, Therapy Planning Using Qualitative Trend Descriptions, in: P. Barahona et al., eds., *Artificial Intelligence in Medicine, Proceedings AIME-95* (Springer, Berlin, 1995) 197–208.

[10] B. Hayes-Roth, R. Washington, D. Ash, R. Hewett, A. Collinot, A. Vina, and A. Seiver, GUARDIAN: A Prototype Intelligent Agent for Intensive-Care Monitoring, *Artif. Intell. Med.* **4** (1992) 165–85.

[11] S. Uckun, B.M. Dawant, E.J. Manders and D.P. Lindstrom, SIMON: An Integrated Approach to Patient Monitoring in Critical Environments, in: K.C. Lun et al., eds., *MEDINFO 92* (North-Holland, Amsterdam, 1992) 564–569.

[12] S. Uckun, B.M. Dawant and D.P. Lindstrom, Model-based Diagnosis in Intensive Care Monitoring: the YAQ Approach, *Artif. Intell. Med.*, **5** (1993) 31–48.

[13] M. Dojat and C. Sayettat, Aggregation and Forgetting: Two Key Mechanisms for Across-Time Reasoning in Patient Monitoring, in I.S. Kohane, et al., eds., *AI in Medicine: Interpreting Clinical Data* (Working Notes AAAI Spring Symposium Series, AAAI Press, Menlo Park, 1994) 33–36.

[14] M. Dojat and C. Sayettat, A Realistic Model for Temporal Reasoning in Real-time Patient Monitoring, *Appl. Artif. Intell.* **10** (1986) 121–143.

[15] Y. Shahar and M.A. Musen, RÉSUMÉ: A Temporal-Abstraction System for Patient Monitoring, *Comput. Biomed. Res.* **26** (1993) 255–273.

[16] Y. Shahar and M.A. Musen, Knowledge-Based Temporal Abstraction in Clinical Domains, *Artif. Intell. Med.* **8** (1996) 267–298.

[17] S. Miksch, W. Horn, C. Popow and F. Paky, Context-Sensitive and Expectation-Guided Temporal Abstraction of High-Frequency Data, in: Y. Iwasaki and A. Farquhar, eds., *Qualitative Reasoning: The Tenth International Workshop* (AAAI Press, Menlo Park, TR WS-96-01, 1996) 154–163.

[18] D. Marsden, M.C. Chiu, F. Paky and P. Helms, Transcutaneous Oxygen and Carbon Dioxide Monitoring in Intensive Care, *Arch. Dis. Childhood* **60** (1985) 1158–1161.

[19] J. Højstrup, A Statistical Data Screening Procedure, *Measurem. Sc. and Technol.* **4** (1992) 153–157.

[20] G. Egghart, Validierung von In- und Outputdaten: Methoden und ihre Anwendungen in VIE-VENT, Master Thesis (Department of Medical Cybernetics and Artificial Intelligence, University of Vienna, 1995).

[21] I.J. Haimowitz, P.P. Le and I.S. Kohane, Clinical Monitoring Using Regression-Based Trend Templates, *Artif. Intell. Med.* **7** (1995) 473–496.

[22] T.A. Russ, Use of Data Abstraction Methods to Simplify Monitoring, *Artif. Intell. Med.* **7** (1995) 497–451.

[23] C. Larizza, A. Moglia and M. Stefanelli, M-HTP: A System for Monitoring Heart Transplant Patients, *Artif. Intell. Med.* **4** (1992) 111–126.

[24] L. Chittaro, M. del Rosso and M. Dojat, Modeling Medical Reasoning with the Event Calculus: An Application to the Management of Mechanical Ventilation, in P. Barahona, et al., eds., *Artificial Intelligence in Medicine, Proceedings AIME-95* (Springer, Berlin, 1995) 79–90.

[25] D.F. Sittig and M. Factor, Physiologic Trend Detection and Artificial Rejection: a Parallel Implementation of Multi-State Kalman Filtering Algorithm, *Comp. Meth. Progr. Biomed.* **31** (1990) 1–10.

[26] D.F. Sittig and J.A. Orr, Evaluation of a Parallel Implementation of the Learning Portion of the Backward Error Propagation Neural Network: Experiments in Artifact Identification, in: P.D. Clayton, ed., *Proceedings of the Fifteenth Annual Symposium on Computer Applications in Medical Care, SCAMC-91* (McGraw-Hill, New York, 1992) 290–294.

[27] J.J. van der Aa, Intelligent Alarms in Anesthesia. Thesis (Technische Universiteit Eindhoven, The Netherlands, 1990).

[28] T. Sukuvaara, E.M.J. Koski, A. Maekivirta and A. Kari, A Knowledge-Based Alarm System for Monitoring Cardiac Operated Patients - Technical Construction and Evaluation, *Int. J. Clin. Mon. Comp.* **10** (1992) 117–126.

[29] D. Garfinkel, P.V. Matsiras, T. Mavrides, J. McAdems and S.J. Aukburg, Patient Monitoring in the Operating Room: Validation of Instruments Reading by Artificial Intelligence Methods, in: L.C. Kingsland, ed., *Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care, SCAMC-89* (IEEE Computer Society Press, Washington, DC, 1989) 575–579.

[30] A. Riva and R. Bellazzi, Learning Temporal Probabilistic Causal Models from Longitudinal Data, *Artif. Intell. Med.* **8** (1996) 217–234.

[31] P. Dagum and A. Galper, Time Series Prediction Using Belief Network Models, *Int. J. Human-Comp. Stud.* **42** (1995) 617–632.

About the Authors

Werner Horn

Werner Horn, Ph.D., is Associate Professor of Artificial Intelligence at the Department of Medical Cybernetics and Artificial Intelligence of the University

of Vienna. He studied Computer Science at the Vienna University of Technology. Since 1984 he has been head of the Knowledge-Based Systems Group of the Austrian Research Institute for Artificial Intelligence. From 1992 onwards he acted as a Representative in the European Coordinating Committee on Artificial Intelligence.

His main research interest include knowledge modeling, expert systems, knowledge acquisition, knowledge engineering, and specifically the embedding of knowledge-based modules into medical application systems.

Silvia Miksch

Silvia Miksch, Ph.D., studied Economic Computer Science at the University of Vienna and the Vienna University of Technology with concentration on: Applied Statistics and Artificial Intelligence. She received her master of Social and Economic Science from the University of Vienna in 1987 and her Ph.D. in 1990. Afterwards she joined the Knowledge-Based Systems Group at the Austrian Research Institute for Artificial Intelligence (OFAI) as scientific researcher and was visiting scholar at Knowledge System Laboratory (KSL), Stanford University, CA, USA (FWF-Erwin Schrödinger post-graduate fellowship). Since 1996, she is Assistant Professor at the Department of Software Technology, Vienna University of Technology.

Her general research interests are: temporal representation and reasoning, task-oriented design, protocol- and guideline-based care, planning, scheduling, time series analysis and filtering techniques, visualization, and evaluation of knowledge-based systems in real-world environments.

Gerhilde Egghart

Gerhilde Egghart studied Computer Science at the Vienna University of Technology. Her field of specialization was artificial intelligence and theoretical computer science. She finished her studies in 1995 with a diploma thesis on validation methods and their application in VIE-VENT. She is presently working for a consultation company on the design and development of tools for marketing and sales.

Christian Popow, MD, was born in Vienna, Austria, 1950. He is presently working as associate professor of Pediatrics and vice head of the Department of Neonatology and Intensive Care of the Department of Pediatrics, Faculty of Medicine, University of Vienna, Austria. He studied economics and medicine in Vienna, Austria, specialized in Pediatrics and Adolescent Medicine, subspeciality Neonatology.

His special fields of interest include clinical neonatology, lung mechanics, mechanical ventilation of newborn infants, especially high frequency oscillatory ventilation; applied computer science, especially patient data management systems and medical expert systems.

Franz Paky

Frank Paky, MD, is head of the Department of Pediatrics of the Hospital of Mödling in Austria. He is practicing pediatrician in Vienna. He studied medicine at the University of Vienna Medical School. In 1984 he was member of the Respiratory Intensive Care Unit of the Hospital for Sick Children in London. From 1984 until 1993 he was heading the Neonatology and Pediatric Intensive Care Unit of the Preyer'sche Children Hospital in Vienna.