Monitoring and Therapy Planning without Effective Data Validation are Ineffective

Silvia Miksch[1*)], Werner Horn[1,2)], Gerhilde Egghart[2)]
Christian Popow[3)], Franz Paky[4)]
[1)] Austrian Research Institute for Artificial Intelligence
Schottengasse 3, A-1010 Vienna, Austria
[2)] Department of Medical Cybernetics and Artificial Intelligence, University of Vienna
[3)] NICU, Division of Neonatology, Department of Pediatrics, University of Vienna
[4)] Department of Pediatrics, Hospital of Mödling, Austria
Email: silvia@ai.univie.ac.at

January 24, 1996

Abstract

Systems for monitoring and therapy planning, which receive their data from computer-based patient records and on-line monitoring equipment, require reliable data. Reasoning on faulty data can cause unexplainable and life-threatening conclusions. Effective and efficient data validation methods are needed to arrive at reliable conclusions.

We distinguished four categories of data validation and repair based on their underlying temporal ontologies: time-point-, time-interval-, trend-based, and time-independent validation and repair. Observing single measurements is not effective to arrive at trustable data. Therefore we take into account the behavior of parameters in the past as well as knowledge derived from domain experts. Examples from VIE-VENT, a knowledge-based monitoring and therapy-planning system for artificially-ventilated newborns, demonstrate the applicability of these methods.

1 Introduction

In the last years, several sophisticated knowledge-based monitoring and therapy-planning systems have been introduced. These systems concentrated on optimizing data analyses and interpretation based on temporal abstraction mechanisms [Shahar and Musen 96, Haimowitz et al. 95, Miksch et al. 95], on applying different kinds of accessible knowledge and information to enrich the reasoning process [Hayes-Roth et al. 92], 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 (CPR). Nevertheless, effective and efficient monitoring and therapy planning require reliable data [Carlson et al. 95]. Real data are more faulty than is often realized. The importance of data validation has been neglected in the past.

We evaluated on-line data sets obtained from newborn infants with various respiratory illnesses. The data (mostly transcutaneous blood gas measurements and oxygen saturation) were collected from the monitoring system of a neonatal intensive care unit (NICU) once per second

*currently visiting scholar at Knowledge Systems Lab, Stanford University. 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, Research, and the Arts, Vienna.

(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, and 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).

In the following sections we will discuss the need for effective data validation, show the approach taken in our monitoring and therapy-planning system VIE-VENT, and present the methods used.

2 The Need for Effective Data Validation

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 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. 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. Avoiding false and providing 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

These findings led us to concentrate on the data validation component in our system VIE-VENT [Miksch et al. 93], an open-loop, knowledge-based monitoring and therapy-planning system for artificially-ventilated newborn infants. We performed a two-step data-validation process based on various kinds of real 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.

4 Data Validation Methods

We distinguished four categories of data validation and repair based on their underlying temporal ontologies: time-point-, time-interval-, trend-based, and time-independent validation and repair. In the following, the order of listing the methods expresses the order of their application. An introductory discussion of the basic methods for data validation including an explanation of the time-point-based data abstraction methods is given in [Miksch et al. 94].

4.1 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 for the reasoning process. 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 is basically simple but has shown very powerful to detect disconnections and missing measurements. We have enhanced the method by adding a clinical context. This context modifies the range for plausible values depending on the mode of ventilation.

Causal dependencies allow to establish a relationship between different parameters. Qualitative values (e.g., $chest\,wall\,extension = small$) are related to numerical ranges of other parameters (e.g., $tidalvolume \leq 5ml/kg$). A causal dependency may 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 ($PULS$) or if we detect a substantial difference between the pulse and the heart rate from ECG (HR, measured in $beats/min$):

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

$$\mid HR - PULS \mid > 8 \rightarrow valid(S_aO_2) = false \tag{2}$$

This example demonstrates the complexity in scheduling the validation process. First we have to use all known methods to validate PULS and HR. If both are valid we apply equation (2).

Functional 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.,

$$f = \frac{60}{t_i + t_e} \tag{3}$$

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 the right side of the equation with the maximum allowed difference due to the explained error of the parameters.

Functional qualitative dependencies establish a relationship between derived qualitative values of different parameters. 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. VIE-VENT uses an unified scheme for all blood gas measurements:

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

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 4.4 and the dynamic reliability score computed by each of the various validation methods. 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 time t_x with the results available after some minutes, say at time 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 may 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.

4.2 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 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 is set invalid, 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.

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

Allowed changes of values of parameters is the comparison of 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: this is a strong method to perform a stability check. After a period of invalidity 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 first value of parameter, which is classified valid by all

other validation methods becomes a candidate for stability testing. During time interval n we require, e.g.,

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

For excellent stability of S_aO_2 we currently use $n = 120sec$ and $\varepsilon = 5\%$.

- 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 of data from different sources is the time-interval-based utilization of functional dependencies described in section 4.1. 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_cCO_2 and $P_{tc}CO_2$ as described above, and both parameters are not invalidated by other methods, we directly apply 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.

Based on 442 cases with corresponding measurements we were able to find a correlation function between P_aCO_2 and $P_{tc}CO_2$:

$$P_{tc}CO_2^{corr} = 2.226 + 1.039 P_aCO_2, r^2 = 0.705 \tag{5}$$

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 $t_y = t_{x+m}$:

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

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

4.3 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), which overcome the limitations of predefined static thresholds. We applied range checks on the growth rate, an evaluation procedure, which inspects the temporal behavior of measurements (Højstrup method modified), trend-based functional dependencies of different dependent variables, and an assessment procedure of the development of a variable. There are two possibilities to handle missing or invalid 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 a linear regression model based on the short-term trend (i.e., of the last 10 minutes) to predict a "correct" value.

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 the 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 may easily be adapted to a specific clinical situation based on the frequency at which data values arrive. The following table summarizes the trends and their criteria:

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%

For each kind of trend the actual growth rate k and the derived qualitative trend category is determined.

Range checking of the growth rate is a sensible method for recognizing problems with the technical equipment, e.g., sensor loss. They are applied on the very short-term trend and react thus very fast.

The Højstrup method modified recognizes growth rates unacceptable after a certain amount of time. The basic idea is given in [Højstrup 92]. It predicts a next data value v_i using the last value received x_{i-1}, the mean value M of a sequence of data points, and the correlation K of two neighboring points:

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

After getting the new value x_i it updates M and K. If the difference between v_i and x_i exceeds a threshold x_i is invalid. Based on the assumption that this difference follows a Gaussian distribution we fix the error threshold E to a value that the probability is less than, e.g., 0.01 for a correct value exceeding 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 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 works as follows:

1. Using the last measurement x_{i-1}, mean m_{i-1}, and standard deviation s_{i-1} predict the next value v_i:

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

2. Get the new data value x_i

3. Update mean and standard deviation:

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

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

4. Decide whether x_i is valid:

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

5. Continue with next i.

The critical part of the algorithm is the fine tuning of the determining parameters M, R, and E. Based on a systematic analysis we are able to derive M, R, and E from three plausible parameters: the sampling rate, the steepest valid growth, and the rigidity. Several experiments have been done to get the desired behavior for each of the continuously–assessed parameters.

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.

Expected qualitative trend descriptions 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 $P_{tc}O_2$ is moving one qualitative step towards the target range within 20 to 30 minutes". Applying these qualitative trend descriptions to the data-point categories defined in section 4.1 we get a qualitative notion of "normal decrease", which is defined by an area of (negative) growth. The supposed exponential functions, which delimit this area, 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 value into one of ten qualitative trend categories. The categories are divided by the target range into an upper and a lower region:

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

Details about the trend-curve-fitting scheme are to be found in [Miksch et al. 95].

Trend-based functional dependencies model expectations on trends. They use the qualitative trend categories to compare the trends of related parameters, e.g., the trend of S_aO_2 and $P_{tc}O_2$. The comparison has to be done using the short-term trend and the medium-term trend. If trends differ by more than one category both measurements are marked ambiguous.

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 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 schema and its resulting qualitative trend categories dynamically models this behavior.

4.4 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 allows to identify 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. On the one hand these lists facilitate the data validation task in case of bidirectional dependencies. On the other hand they also help 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.

5 Conclusion

Applying our validation methods to the observed on- 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 minimized, (2) errors of data interpretation were reduced, (3) abrupt changes of therapeutic recommendations were eliminated promoting a stable and graceful weaning process.

References

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

[Haimowitz et al. 95] Haimowitz I.J., Le P.P., Kohane I.S.: Clinical Monitoring Using Regression-Based Trend Templates. *Artificial Intelligence in Medicine*, 7(6):473–496, 1995.

[Hayes-Roth et al. 92] Hayes-Roth B., Washington R., Ash D., Hewett R., Collinot A., Vina A., Seiver A.: GUARDIAN: A Prototype Intelligent Agent for Intensive-Care Monitoring. *Artificial Intelligence in Medicine*, 4(2):165–185, 1992.

[Højstrup 92] Højstrup J.: A Statistical Data Screening Procedure. *Measurement Science and Technology*, 4:153–157, 1992.

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

[Miksch et al. 94] Miksch S., Horn W., Popow C., Paky F.: Context-Sensitive Data Validation and Data Abstraction for Knowledge-Based Monitoring. In Cohn A.G.(ed.), *Proc. 11th European Conference on Artificial Intelligence (ECAI94)*. Wiley, Chichester, UK, pp.48–52, 1994.

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

[Shahar and Musen 96] Shahar Y., Musen M.A.: Knowledge-Based Temporal Abstraction in Clinical Domains. *Artificial Intelligence in Medicine, Special Issue "Temporal Reasoning in Medicine"*, (forthcoming) 1996.