

## Mining Temporal Data Sets: Hypoplastic Left Heart Syndrome Case Study

Andrew Kusiak\*, Christopher A. Caldarone\*\*, Michael D. Kelleher\*\*\*,  
Fred S. Lamb\*\*\*, Thomas J. Persoon\*, Yuan Gan\*, and Alex Burns\*

\*Intelligent Systems Laboratory, 2139 Seamans Center

\*\*Department of Cardiothoracic Surgery and \*\*\*Pediatrics

University of Iowa Hospital and Clinics

The University of Iowa

Iowa City, Iowa 52242 - 1527

Tel: 319 - 335 5934 Fax: 319 - 335 5669

andrew-kusiak@uiowa.edu

<http://www.icaen.uiowa.edu/~ankusiak>

### ABSTRACT

Hypoplastic left heart syndrome (HLHS) affects infants and is uniformly fatal without surgery. Post-surgery mortality rates are highly variable and dependent on postoperative management. The high mortality after the first stage surgery usually occurs within the first few days after procedure. Typically, the deaths are attributed to the unstable balance between the pulmonary and systemic circulations. An experienced team of physicians, nurses, and therapists is required to successfully manage the infant. However, even the most experienced teams report significant mortality due to the extremely complex relationships among physiologic parameters in a given patient. A data acquisition system was developed for the simultaneous collection of 73 physiologic, laboratory, and nurse-assessed variables. Data records were created at intervals of 30 seconds. An expert-validated wellness score was computed for each data record. A training data set consisting of over 5000 data records from multiple patients was collected. Preliminary results demonstrated that the knowledge discovery approach was over 94.57% accurate in predicting the "wellness score" of an infant. The discovered knowledge can improve care of complex patients by the development of an intelligent simulator that can be used to support decisions.

Keywords: Temporal data mining, hypoplastic left heart syndrome, medical informatics, derived features, feature transformation, data mining expressions.

## 1. INTRODUCTION

In this paper, mining of temporal data sets is considered. The concept of horizontal and vertical expressions is illustrated on the data set created during the postoperative care of neonates after the first stage of the Norwood procedure. This stage involves neonatal reconstruction of the heart soon after birth. Although the procedure is lifesaving, a consequence of this surgery is a precarious unstable balance between the pulmonary and systemic circulation. Unfortunately, the mortality after this first stage surgery can reach 42% (CHA 2002). In this paper, machine learning algorithms are used to discover knowledge about the complex relationships between the large numbers of physiological variables and a health status of an infant. The discovered knowledge is used to improve understanding of the decisions during neonate postoperative management and to predict the effectiveness of intended interventions.

There are two accepted therapeutic strategies to manage a neonate with hypoplastic left heart syndrome (HLHS), the Norwood procedure and cardiac transplantation (Ohye 2002). The Norwood procedure has emerged in the last decade as the most common treatment (Gutgesell 2002) and it involves a three stage surgical intervention (Ohye 2002). Stage I of the Norwood procedure includes three main components: an atrial septectomy, an anastomosis of the proximal pulmonary artery to the aorta with homograft augmentation of the aortic arch, and an aortopulmonary shunt. As a result of this procedure the patient's right ventricle is connected to the aorta so that it can force the delivery of oxygenated blood through the branches of the aorta.

The most critical time for the neonate is the surgery itself and the time immediately following surgery spent in the Pediatric Intensive Care Unit (PICU). Typically, complications are attributed to the unstable balance between the pulmonary and systemic circulation. There are rapid and massive shifts in the cardiac output, pulmonary resistance, and systemic resistance for the first 3-4 days after surgery. An experienced team of physicians, nurses, and therapists is required during this period of time. However, even the most experienced teams report significant mortality due to the extremely complex relationships among physiologic parameters in a given patient. The mortality rate for the three-stage procedure is highest following the first stage (Tulloh 2001). This high rate is due the inability to directly measure crucial parameters in the post-operative infant. The physicians need to infer the value of crucial, but immeasurable, parameters from a group of obtainable parameters used to monitor the infant.

Obtainable postoperative parameters include: heart rate, heart rhythm, systemic blood pressure, common atrial filling pressure, urine output, physical exam, and systemic and mixed venous oxygen saturations. Based on these values, inferences are made as to the value of crucial life-saving parameters (e.g., pulmonary and systemic blood flow). These parameters change rapidly in the postoperative period, and subtle constellations of changes in the obtainable parameters are often unnoticed by the inexperienced caregiver but lead to a "sudden" postoperative death. Closer analysis of the medical record often reveals premonitory clusters of changes that should have signaled a modification of the direction of postoperative therapy.

## 2. RESEARCH METHODOLOGY

### 2.1 Expressions of Temporal Patterns

Consider the temporal data set in Table 1 with four features collected every second for the period of 12 seconds.

The following three experiments are performed with this data set in Table 1:

- Experiment 1: Rules are extracted from the data set with all features included (see Figure 1).
- Experiment 2: Rules are extracted from the data set with two features only; Features\_1 and Feature\_2 (see Figure 2).
- Experiment 3: Rules are extracted from the data set with two features only; Features\_3 and Feature\_4 (see Figure 3).

Table 1. Sample data set.

Time	Feature_1	Feature_2	Feature_3	Feature_4	Outcome
1	1.1	14.1	a	p	A
2	2.1	14.1	b	q	A
3	3.1	14.1	a	p	A
4	3.1	12.3	d	r	B
5	3.1	11.2	d	r	B
6	3.1	10.1	c	s	B
7	2.9	17.4	c	s	B
8	2.7	18.3	d	r	B
9	2.5	19.2	c	s	B
10	2.3	16.1	b	q	A
11	2.3	13.0	b	q	A
12	2.3	9.90	b	q	A

- Rule 1. IF (Feature\_1 < 2.4) THEN (Outcome = A); [5, 5, 100.00%]  
[1, 2, 10, 11, 12]
- Rule 2. IF (Feature\_3 = a) THEN (Outcome = A); [2, 2, 100.00%]  
[1, 3]
- Rule 3. IF (Feature\_1 >= 2.4) AND (Feature\_3 in {d, c}) THEN (Outcome = B); [6, 6, 100.00%][4, 5, 6, 7, 8, 9]

Figure 1. Decision rules extracted from the data set in Table 1 with all four features included.

The decision rules are presented in the following format:

IF (Conditions) THEN (Outcome); [Rule support], [Rule strength], [Rule confidence], [Objects represented by the rule]

- Rule 1. IF (Feature\_1 < 2.4) THEN (Outcome = A); [5, 5, 100.00%]  
[1, 2, 10, 11, 12]
- Rule 2. IF (Feature\_1 >= 3) AND (Feature\_3 = a) THEN (Outcome = A); [1, 1, 100.00%][3]
- Rule 3. IF (Feature\_1 in [2.4, 3)) THEN (Outcome = B); [3, 3, 100.00%][7, 8, 9]
- Rule 4. IF (Feature\_1 >= 2.4) AND (Feature\_2 < 12.65) THEN (Outcome = B); [3, 3, 100.00%][4, 5, 6]

Figure 2. Decision rules extracted from the data set in Table 1 with Features\_1 and Feature\_2 included.

- Rule 1. IF (Feature\_4 in {q, p}) THEN (Outcome = A); [5, 5, 100.00%]  
[1, 2, 3, 10, 12]
- Rule 2. IF (Feature\_3 = b) THEN (Outcome = A); [1, 1, 100.00%][11]
- Rule 3. IF (Feature\_3 in {d, c}) AND (Feature\_4 in {s, r}) THEN (Outcome = B);  
[6, 6, 100.00%][4, 5, 6, 7, 8, 9]

Figure 3. Decision rules extracted from the data set in Table 1 with Features\_3 and Feature\_4 included.

A. Kusiak, C.A. Caldarone, M.D. Kelleher, F.S. Lamb, T.J. Persoon, Y. Gan, and A. Burns, Mining Temporal Data Sets: Hypoplastic Left Heart Syndrome Case Study, in B.V. Dasarathy (Ed.), *Proceedings of the SPIE Conference on Data Mining and Knowledge Discovery: Theory, Tools, and Technology V*, SPIE, Orlando, FL, April 2003, pp. 93-101.

The temporal nature of the data set in Table 1 is expressed vertically (column wise) and horizontally (row wise). These temporal data expressions can be represented and analyzed in many different ways. Horizontal feature expressions (e.g., feature sequences) are discussed in Kusiak (2001). One of the simplest ways of representing the vertical temporal expression is with a feature – time function. Such a function for Feature\_1 of Table 1 is shown in Figure 4.

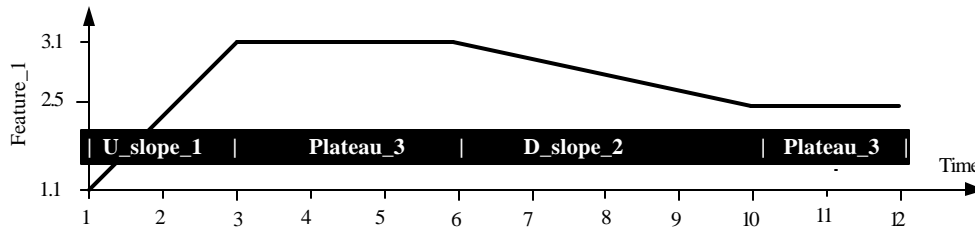


Figure 4. Plot of Feature\_1 in time.

The function in Figure 4 has been partitioned into four regions labeled as U\_slope\_1, Plateau\_3, D\_slope\_2, and Plateau\_2. Each of the four partitions represents a clearly recognizable pattern and has the same decision (Outcome in Table 1) value. The latter opens up an issue whether a feature function in time or decision should be considered. In addition to plotting single features, functions of more than one feature could be considered.

The four regions labeled in Figure 4 as U\_slope\_1, Plateau\_3, D\_slope\_2, and Plateau\_2 make up the values of a new feature, the derived feature D\_Feature\_1. Though derived features have been used in data mining, this paper introduces a way of deriving such features by using feature expressions.

Similarly to the derived feature D\_Feature\_1, the derived feature D\_Feature\_2 has been created (see Table 2).

Table 2. The data set of Table 1 with three derived features D\_Feature\_1, \_Feature\_2, D\_Feature\_3\_4.

Time	F_1	D_Feature_1	F_2	D_Feature_2	F_3	F_4	D_Feature_3_4	Outcome
1	1.1	U_slope_1	14.1	Plateau_14	a	p	a_p	A
2	2.1	U_slope_1	14.1	Plateau_14	c	q	c_q	A
3	3.1	U_slope_1	14.1	Plateau_14	a	p	a_p	A
4	3.1	Plateau_3	12.3	D_slope_12	d	r	d_r	B
5	3.1	Plateau_3	11.2	D_slope_12	d	r	d_r	B
6	3.1	Plateau_3	10.1	D_slope_12	c	s	c_s	B
7	2.9	D_slope_2	17.4	U_slope_17	c	s	c_s	B
8	2.7	D_slope_2	18.3	U_slope_17	d	r	d_r	B
9	2.5	D_slope_2	19.2	U_slope_17	c	s	c_s	B
10	2.3	Plateau_2	16.1	D_slope_16	c	q	c_q	A
11	2.3	Plateau_2	13	D_slope_16	b	r	b_r	A
12	2.3	Plateau_2	9.9	D_slope_16	c	q	c_q	A

The derived features D\_Feature\_1 and D\_Feature\_2 have been created through the vertical expression (feature function plot) of temporal features. The horizontal expression of two features Feature\_3 and Feature\_4 of Table 1 is shown in Table 2 as the derived feature D\_Feature\_3\_4. This feature was created by combining the corresponding values of the two original features: Feature\_3 and Feature\_4 (denoted for short in Table 2 as F\_3 and F\_4).

The impact of the derived features on the decision rules is discussed next. First, consider the rules extracted from the data set in Table 2 with two derived features D\_Feature\_1 and D\_Feature\_2 only.

```
Rule 1. IF (D_Feature_1 in {U_slope_1, Plateau_2}) THEN (Outcome = A); [6, 6,
    100.00%][1, 2, 3, 10, 11, 12]
Rule 2. IF (D_Feature_1 in {Plateau_3, D_slope_2}) THEN (Outcome = B); [6, 6,
    100.00%][4, 5, 6, 7, 8, 9]
```

Figure 5. Decision rules extracted from the data set in Table 2 with D\_Features\_1 and D\_Feature\_2 included.

Compared to the rule set in Figure 2 derived from the dataset with the two original features Features\_1 and Feature\_2, the rule set in Figure 5 contains 50% of the number of original rules. Though the two data sets are small, it can be shown that the classification accuracy of the rules extracted based on the derived features is higher than that of original features. In this case the classification accuracy is 91.67% (derived features) vs 83.33% (original features) using on the one-out-of-eight cross-validation scheme.

The rule set extracted from the data set in Table 2 with the derived feature D\_Feature\_3\_4 is shown in Figure 6.

```
Rule 1. IF (D_Feature_3_4 in {c_q, a_p, b_r}) THEN (Outcome = A); [6, 6,
    100.00%][1, 2, 3, 10, 11, 12]
Rule 2. IF (D_Feature_3_4 in {d_r, c_s}) THEN (Outcome = B); [6, 6, 100.00%]
    [4, 5, 6, 7, 8, 9]
```

Figure 6. Decision rules extracted from the data set in Table 2 with the derived feature D\_Features\_3\_4 included.

Again, this rule set is smaller than the corresponding one in Figure 3.

The decision rules extracted from data sets with derived features based on the vertical and horizontal expressions provide numerous advantages, such as:

- Improved generality: Derived features generalize the original features and this may lead to more general knowledge.
- Improved comprehensibility: Rule sets of smaller sizes are normally better understood and accepted by the users.
- User preferences: A user may want to view knowledge with predetermined configuration of features, which can be accomplished by the definition of vertical and horizontal expressions.

## 2.2 Methods for Creation of Derived Features

The concept of horizontal and vertical expressions appears to be a viable method for generating derived features. Besides the illustrated feature – time function and feature – decision value plots other functions and methods can be used, for example:

- Discretization
- Curve fitting
- Process control charts
- Fourier transforms
- Wavelet functions
- Extreme values
- Moving averages
- User defined preferences



Figure 7. Sample screen.

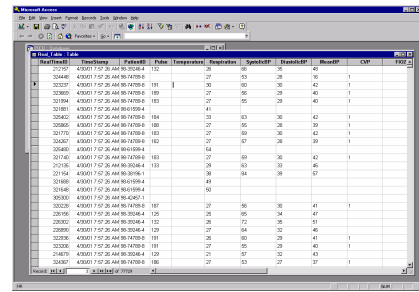


Figure 8. Partial data set.

### 3. MEDICAL CASE STUDY: THE HYPOPLASTIC LEFT HEART SYNDROME TEMPORAL DATA SET

The medical topic considered in this research is complex and it has not been widely covered in the literature. Therefore, the relationships between parameters and the relevancy of parameters monitored in clinical practice to the infant’s health status are not well understood.

Most, if not all, processes in medicine are temporal, which is reflected by the test results and other measurements in time. However, for various reasons they are often considered in the literature as time invariant. To collect the temporal data for the hypoplastic left heart syndrome (HLHS) project, the authors have developed an acquisition system (see sample screens in Figure 7 and Figure 8) that collects data at discrete intervals. The sampling frequency was determined intuitively based on the most dynamic process component. In most cases, prior experience and availability of sensors have dictated the spectrum of the data collected. One of the side benefits of this research is an algorithm for the selection of most promising parameters and a rational sampling frequency.

The research team has been involved in three other application domains similar to the one considered in this research, namely lung cancer diagnosis (Kusiak *et al.* 2000), infants’ heart arrhythmia (Kusiak *et al.* 2001), and patient dialysis. Besides medical applications the research team has been involved in other knowledge discovery projects.

A subset of features and their values leading to a desirable outcome constitutes a decision signature (Kusiak 2002). The proposed research is designed to identify decision signatures and use them to improve care for infants after heart surgery. If there was just a single decision signature associated with a particular outcome, there would be little room for research. Due to temporal nature of the data, the decision signatures are valid for a limited time, which leads to the problem addressed in this paper.

Table 3. Continuously monitored parameters.

Parameter
Pulse
Respiration
Systolic blood pressure
Diastolic blood pressure
Mean blood pressure
Central venous pressure
Oxygen saturation via pulse oximeter

The continuously monitored physiologic parameters are listed in Table 3. Transducers to measure these parameters were attached or implanted in the patient and were connected to a bedside monitoring device. Individual patient monitors transmitted the data to a patient data server (PDS), which collated the data from all patients in the Pediatric Intensive Care Unit (PICU). Collated data was transmitted via Ethernet to a data collection server. Custom software in the server pulled parameters of interest from the data stream and stored them in a database.

The intermittently monitored parameters consisted of physiologic parameters observed at the bedside by nurses and laboratory data. These parameters were obtained retrospectively from nursing assessment records (flow sheets) and are listed in Table 4.

Interventions are actions taken by the caregiver to treat the patients, primarily administration of medications, and are listed in Table 5. Both continuously administered intravenous medications and intermittently administered medications

are included, however substances administered primarily as vehicles for other medications (e.g., normal saline solution) are not included.

The three categories of data, continuous and intermittently monitored physiologic parameters, and interventions, combined for a total of 73 variables that were collected by a computerized data acquisition system developed by the research team. The acquisition system collects data from the PICU equipment and stores it in an Access database. A data

Table 4. Intermittently monitored parameters.

Parameter
Venous oxygen saturation
Hemoglobin concentration
Blood sodium concentration
Blood potassium concentration
Blood chloride concentration
Blood carbon dioxide concentration
Blood ionized calcium concentration
Blood glucose concentration
Blood pH
Blood partial pressure of CO <sub>2</sub>
Blood partial pressure of oxygen
Blood carbon dioxide concentration
Blood base excess
Blood lactic acid concentration
Radial arterial blood pressure
Femoral arterial blood pressure
Color
Presence or absence of bowel sounds
Condition of patient's abdomen
Assessment of patient's breath sounds
Peripheral pulse, left upper extremity
Peripheral pulse, right upper extremity
Peripheral pulse, left lower extremity
Peripheral pulse, right lower extremity
Capillary refill, seconds
Right pupil size
Right pupil reaction to light
Left pupil size
Left pupil reaction to light
Urine output volume
Chest tube drainage volume
Mediastinal tube volume
Total fluid output

Table 5. Interventions.

Intervention
Administration of epinephrine
Administration of norepinephrine
Administration of dopamine
Administration of milrinone
Change ventilator
Administer HCO <sub>3</sub>
Administer FFP
Administer KCl
Administer CaCl
Administer PRBC
Administer Lasix
Administer albumin
Ice to heart
Pacer on/off
Administer amiodarone
Liver pressure
Epi bolus
Change fluid input
Administer platelets
Administer atropine
CPR
Administer insulin
Internal heart compression
Defib
Administer Mg
Nipride

set for five patients was created for the preliminary analysis. The acquisition system can practically collect data on any number of patients for any length of time. The five patient's data amounted to over 5,000 data objects.

A data object is defined as a vector of attribute values. Attributes of a data object include all 73 of the continuous and intermittently assessed parameters collected a specific instance of time and all of the interventions being administered at that instance. For ease of analysis, data objects were transferred from the Access database into a spreadsheet, with the rows representing data objects and the columns representing attributes. The time attribute of each object was recorded in minutes following post surgical admission of a neonate to the PICU.

For each data object, an expert – validated wellness score was determined. The individual elements of the wellness score are listed in Table 6. If a wellness score attribute fell within the “good” range the element was scored as 1; otherwise it was scored as 0. The clinical experts also determined that one intervention, administration of ionotropic drugs, should also be a constituent of the wellness score. If a patient required no ionotropic support, the ionotrope wellness score is 1. The ionotrope wellness score was adjusted by addition of each ionotrope according to the scheme shown in Table 7. The ionotrope portion of the wellness score could range from 1 (no ionotropes) to -3 (four ionotropes)

administered at maximum concentration). The total wellness score is the sum of the individual wellness score components.

The initial mining of the retrospective data set demonstrated that the extracted rules can predict the wellness score with accuracy of 94.57%. The rules made incorrect predictions for 1.60% of new cases (with unknown value of the wellness score) and for 2.22% of new cases the decision rules could not make any predications. Some outcomes, e.g., the wellness score of 7.5 that were represented by sufficient number of objects could be predicted with 100% accuracy.

Table 6. Elements of the wellness score.

Parameter	Range for Wellness Score = 1	Range for Wellness Score = 0
Heart Rate	=150, = 170	<150, >170
CVP (Preload)	$\geq 8, \leq 14$	<8, >14
SaO <sub>2</sub>	$\geq 65, \leq 90$	<65, >90
SvO <sub>2</sub>	$\geq 20$	<20
Base Excess	$\geq (-5)$	<(-5)
Hemoglobin (Hb)	$\geq 12$	<12
pCO <sub>2</sub>	$\geq 35, \leq 50$	<35, >50
Urine Output	>0.5 cc/kg/hr	<0.5 cc/kg/hr
Inotropic support	None	Defined in Table 7

Using the concept of horizontal and vertical expressions, the predication accuracy was improved. Furthermore the majority of prediction errors were made when the wellness score was less than 3.5. This is due to the insufficient number

Table 7. Wellness score for ionotrope support (doses in mg/kg/hr).

Ionotrope	Score = 0	Score = 0.5	Score = 1
Epinephrine dose	0	<0.1	$\geq 0.1$
Dopamine dose	0	<5	$\geq 5$
Milrinone dose	0	<0.75	$\geq 0.75$
Norepinephrine dose	0	<0.1	$\geq 0.1$

of observations (objects) for these values in the training data set. The increase in the number of patients monitored by the data acquisition system will increase the number of observations of the low wellness score values, which should transform into increased predication accuracy.

#### 4. CONCLUSION

In this paper, a data mining approach was applied to postoperative management of infants with the hypoplastic left heart syndrome. To efficiently analyze the temporal data set, a new concept of horizontal and vertical data expressions was introduced. The vertical expression of data leads to derived features expressed as the function of time or decision value.

Mining the hypoplastic left heart syndrome data set with derived features based on the concept of horizontal and vertical expressions has improved classification accuracy.

#### ACKNOWLEDGEMENT

The authors would like to express appreciation to Andrew Glick for organizing some of the data sets used in the study and C.F. Yun for design and coding the data collection system.

## REFERENCES

- CHA (2002), Hypoplastic left Heart Syndrome, Children's Healthcare of Atlanta, [www.choa.org/library/conditions/hlhs\\_norwood.shtml](http://www.choa.org/library/conditions/hlhs_norwood.shtml).
- Gutgeell, H.P. and J. Gibgon (2002) Management of hypoplastic left heart syndrome in the 1990s, *American Journal of Cardiology*, Vol. 89, No. 7, pp. 842-846
- Kusiak, A. (2001), Feature transformation methods in data mining, *IEEE Transactions on Electronics Packaging Manufacturing*, Vol. 24, No. 3, pp. 214-221.
- Kusiak, A. (2002), A data mining approach for generation of control signatures, *ASME Transactions: Journal of Manufacturing Science and Engineering*, Vol. 124, No. 4, pp. 923-926.
- Kusiak, A., Kern, J.A., Kernstine, K.H., and T.L. Tseng (2000), Autonomous decision-making: A data mining approach, *IEEE Transactions on Information Technology in Biomedicine*, Vol. 4, No. 4, pp. 274-284.
- Kusiak, A., Law, I.H., and M.D. Dick (2001), The G-algorithm for extraction of robust decision rules: Children's postoperative intra-atrial arrhythmia case study, *IEEE Transactions on Information Technology in Biomedicine*, Vol. 5, No. 3, pp. 225-235.
- Ohye, R (2002), *eMedicine Journal*, Vol. 3, No. 89, [www.emedicine.com/ped/topic2825.htm](http://www.emedicine.com/ped/topic2825.htm).
- Tulloh, A., Sharland, G., Simpson, J., Rollings, S., Baker, E., Qureshi, S., Rosenthal, E., Austin, C., and D.P. Anderson (2001), Outcome of staged reconstructive surgery for hypoplastic left heart syndrome following antenatal diagnosis, *Archives of Disease in Childhood*, Vol. 85, No. 6, pp. 474-477.

A. Kusiak, C.A. Caldarone, M.D. Kelleher, F.S. Lamb, T.J. Persoon, Y. Gan, and A. Burns, Mining Temporal Data Sets: Hypoplastic Left Heart Syndrome Case Study, in B.V. Dasarathy (Ed.), *Proceedings of the SPIE Conference on Data Mining and Knowledge Discovery: Theory, Tools, and Technology V*, SPIE, Orlando, FL, April 2003, pp. 93-101.