An Epistemology for Clinically Significant Trends

Ira J. Haimowitz
MIT Laboratory for Computer Science
545 Technology Square, Room 414
Cambridge, MA 02139
ira@medg.lcs.mit.edu

Isaac S. Kohane
Children's Hospital, Harvard Medical School
300 Longwood Avenue
Boston, MA 02115
gasp@medg.lcs.mit.edu

Abstract

We have written a computer program called $TrenD_x$ for automated trend detection during process monitoring. The program uses a representation called *trend templates* that define disorders as typical patterns of relevant variables. These patterns consist of a partially ordered set of temporal intervals with uncertain endpoints. Attached to each temporal interval are value constraints on real-valued functions of measurable parameters. As $TrenD_x$ receives measured data of the monitored process, the program creates hypotheses of how the process has varied over time.

We introduce the importance of a distinct trend representation in knowledge-based systems. Then we demonstrate how trend templates may represent trends that occur at fixed times or at unknown times, and their utility for domains that are quantitatively both poorly and well understood. Finally we present experimental results of $TrenD_x$ diagnosing pediatric growth disorders from heights, weights, bone ages, and pubertal data of twenty patients seen at Boston Children's Hospital. [1]

Introduction

Our work is part of the growing body of artificial intelligence (AI) research on diagnostic process monitoring. Specifically, we have written a program that automatically detects *trends:* sequences of time-ordered data that together are clinically significant. These trends may be multivariate, and may consist of several distinct phases. Our trend detection program, called $TrenD_x$, can classify the trend and give a chronology of when the data was in each phase.

In another paper [Haimowitz and Kohane 1993] we defined our *trend template* representation of clinically significant trends, and illustrated our trend diagnosis program *$TrenD_x$* on a single pediatric growth patient. In this paper we demonstrate how trend templates may represent trends that occur at fixed times or at unknown times. We argue for the utility of trend templates for domains that are quantitatively poorly or well understood. Then we present experimental results of a clinical trial where $TrenD_x$ diagnosed pediatric growth patterns of twenty patients at Boston Children's Hospital.

Need for Trend Representation

Diagnostic knowledge based systems are programs that can reason abductively from symptoms in a patient to the disorders that cause them. However, the vast majority of these programs treat symptoms as fixed in time. These symptoms may be boolean, as in "chest pain = true" or one of a series of qualitative categories for a measurable parameter, as in "serum sodium = low."

Such a stationary representation of findings is insufficient for monitoring a process (such as a patient) being monitored over any period of time. A human expert monitoring a process has notions of *trends:* how the measured parameters should vary over time under the current hpothesis.When the measurements vary from the expected, that expert may consider an alternative diagnosis. For a computer program to behave similarly, it must represent the expected trend.

Merely checking laboratory values against a reference interval can lead to ignoring a trend where the parameter is markedly decreasing, increasing, or periodically fluctuating within that range. A prime example of this comes from the domain of pediatric growth, where heights and weights are measured at least once a year and plotted on growth charts of standard deviations (SDs) for each measurement by age of United States children [Hamil et. al 1979]. A height that decreases from the mean for some age to -1 SD two years later is still within a range of "normal" yet may strongly indicate either an endocrinological or nutritional disorder.

In domains like pediatric growth where one cannot construct a predictive causal model, experts still demonstrate knowledge of how measured parameters vary under different diagnoses. This is the motivation behind our representation for trends called trend templates and our trend diagnosis program $TrenD_x$ that uses them.

1. This work has been supported (in part) by NIH grant R01 LM 04493, NICHHD 5T32 HD07277-9, and by a U.S. Office of Naval Research Graduate Fellowship.

Trend Templates

A *trend template* is an archetypal pattern of data variation in a process disorder. Each trend template has a *temporal component* and a *value component.* The temporal component includes *landmark points* and *intervals.* Landmark points represent significant events in the lifetime of the monitored process. They may be uncertain in time, and so are represented with time ranges *(min max)* expressing the minimal and maximal times between them. Intervals represent periods of the process that are significant for diagnosis or therapy. Intervals consist of begin and end points whose times are declared either as:

- offsets of the form *(min max)* from a landmark point, or
- offsets of the form *(min max)* from another interval's begin or end point.

$TrenD_X$ represents time using the Temporal Utility Package (TUP) of [Kohane 1987]. TUP is a temporal reasoning program with both time points and time intervals; interval structures include a begin point and an end point.

The value component of a trend template is a set of *value constraints* bound to each interval. Each value constraint states that some function of a set of measurable parameters must fall within a certain range. Thus each value constraint is an expression of the form

$$m \leq f(D) \leq M \qquad \text{(EQ 1)}$$

where f is some real valued function defined on patient data, m is a minimum (possibly $-\infty$), and M is a maximum (possibly $+\infty$). In the diagnostic program $TrenD_X$, the function f is evaluated on the set D of multi-parameter data currently assigned to that interval and the result is compared to the bounds m and M.

Another aspect of trend templates models failure-driven triggering of alternate diagnoses. Trend templates include a function *TRIGGER* that computes a set of alternative trend templates for each value constraint and the direction of failure.

$$TRIGGER: vc \times \{low,high\} \rightarrow \{TT_1,TT_2, ..., TT_k\} \qquad \text{(EQ 2)}$$

This function prunes the diagnosis space by localizing small sets of alternate diagnoses to specific temporal intervals and failures.

Trend Template for Normal Growth

An example trend template (Figure 1) expresses the constraints of male average prepubertal growth. Time constraints, expressed in the years of age of a child, are drawn horizontally, and value constraints on real variables are drawn vertically. There are three landmark points: *Birth* occurs at age zero, *Puberty Onset* occurs between ages ten and fifteen years, and *Growth Stops* occurs between ages seventeen and nineteen years. The temporal

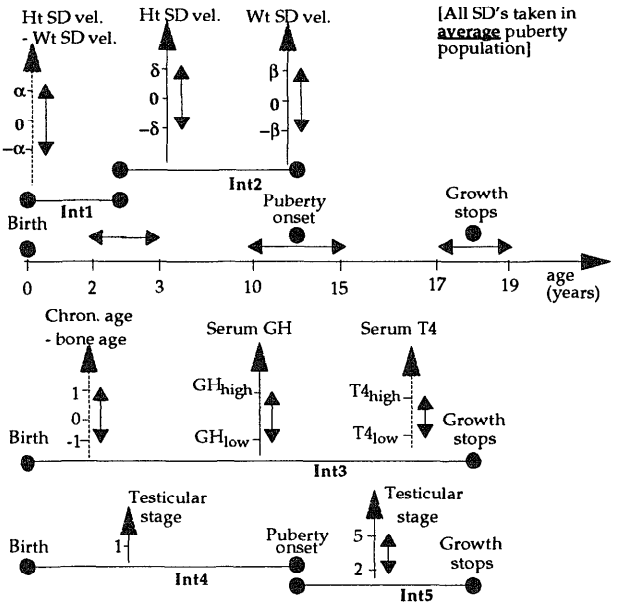

Figure 1 Trend template for male average pre-pubertal growth. Ht is height, Wt is weight, GH is growth hormone, T4 is thyroid hormone, and SD is standard deviation.

uncertainty in these points is depicted with horizontal arrows that span the possible time range.

This trend template contains five intervals. Interval Int1 denotes the time when height and weight standard deviations are established. Int1 begins at *Birth* and ends between ages two and three years. We encode that height and weight standard deviations (SDs) vary in the same way by constraining the difference between the average velocity of height SDs and the average velocity of weight SDs to be within a small number α of zero. Interval Int2 represents the period of the boy staying in his established height and weight channels. Int2 begins at the endpoint of Int1, and Int2 ends at *Puberty Onset.* There are two value constraints: both the average velocities of height SDs and that of weight SDs are close to zero. Because Int1 and Int2 represent consecutive processes of growth, these intervals meet, and we represent the end point of Int1 equal in time to the begin point of Int2. Intervals Int3, Int4 and Int5 constrain other patient parameters: serum growth hormone (GH), serum thyroid hormone (T4), bone age, testicular stage, and other screening tests.

Trigger sets are bound to value constraints of several intervals. For example, if the height SD constraint of Int2 fails low, then the trend template for delayed puberty (termed as "constitutional delay") is suggested. If the constraint fails high, then the trend template for advanced puberty is suggested.

Diagnosis with Trend Templates: TrenD$_X$

The program TrenD$_X$ diagnoses trends by matching process data to trend templates. A TrenD$_X$ hypothesis includes a trend template, an assignment of patient data to the intervals of the template, and a set of temporal assertions that further constrain the endpoints of the template's intervals. TrenD$_X$ initializes a hypothesis for a patient by *anchoring* a landmark point of the trend template as equal to some time in the life of the process. For example,TrenD$_X$ assigns a patient the average normal growth hypothesis by anchoring the *Birth* landmark point of the trend template to the birth date of the patient.

The algorithms for matching a datum *d* to a hypothesis *hyp* are detailed in [Haimowitz and Kohane 1993]. In brief, if *d* meets all value constraints on all intervals that must temporally include that datum, then *hyp* is retained. TrenD$_X$ assigns *d* to all intervals that may contain it. If there are multiple intervals that may include the datum, the program branches to consider distinct data assignments for that same trend template. Because each interval represents a significant process stage, the distinct assignments in fact correspond to alternate hypotheses.

If *d* fails some value constraint on an interval to which it must temporally belong, then *hyp* is removed. The failed value constraint produces a set S of potential new disorders to trigger. The disorders in S are triggered if and only if there exist no other active hypotheses with trend template equal to that of *hyp*. Thus as long as there is some data assignment that is valid for a disorder, TrenD$_X$ does not trigger another disorder.

TrenD$_X$ activates new disorders with a generate and test paradigm. TrenD$_X$ generates the set S of trend templates with the function *TRIGGER*, and tests those templates for matching the patient data. A triggered template becomes an active hypothesis if and only if it matches the patient data.

As TrenD$_X$ monitors a patient, the number of hypotheses increases exponentially in the number N of patient data that may be assigned to multiple intervals of the same trend template. N will be large only if the sampling period of the data is small relative to the uncertainty in the time of interval endpoints. In our experience testing pediatric growth patients with TrenD$_X$ we have always had $N \leq 2$. Furthermore hypotheses have been repeatedly pruned when later data fail some value constraint of a hypothesis with a spurious data assignment.

Expressiveness of Trend Templates

Trend templates are capable of representing trends whether or not the onset time is known beforehand. They may also represent trends in areas of either scant or detailed quantitative models.

Intervening Trends

Thus far we have illustrated a trend template for a pattern where one knows the onset time (birth) of the trend beforehand. Statistical curve-fitting models of pediatric growth [Thissen and Bock 1990] are limited in describing only patterns beginning from birth. However, some trends in process monitoring can appear at any unanticipated time, perhaps due to an unexpected event or as the result of a stimulus not modeled. We call these *intervening trends*. In medicine intervening trends often signify a new disorder.

Pediatric growth disorders marked by intervening trends include nutritional disorders such as malnutrition or obesity. Also included are endocrine disorders such as *acquired growth hormone deficiency*. In this disorder serum growth hormone levels unexpectedly decrease, with a consequent decrease in rate of bone elongation and height. On the growth chart, the child loses height standard deviations, even on standards for patients with delayed puberty. The child also appears heavier over time, which can be detected with an by increase in the body mass index (BMI), calculated as weight/(height)^2, expressed in kg/m^2.

We represent an intervening trend such as in acquired growth hormone deficiency with a trend template that includes a landmark point for the uncertain onset time. Anchoring this landmark point to the patient history requires additional reasoning. TrenD$_X$ must shift the landmark point back in time until all subsequent patient data meets the constraints of the intervening trend's template. Note that because intervening trends represent unanticipated faults, the corresponding template is triggered due to the failure of a value constraint of another hypothesis. TrenD$_X$ must anchor the landmark point (onset time) of the intervening trend's template earlier in time than the data failing the value constraint.

Below is the trend template for acquired growth hormone deficiency:

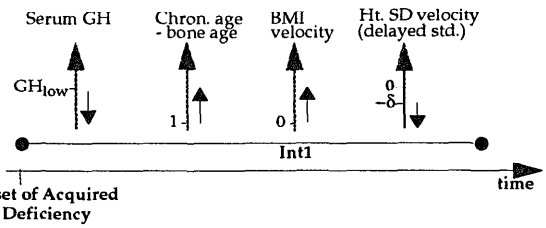

Figure 2 Trend template for acquired growth hormone deficiency. BMI is body mass index.

The landmark point for this trend template denotes the onset time of the growth hormone deficiency. Int1 is an

interval beginning at that time and contains four value constraints: serum GH is less than the lower threshold mentioned in Figure 1, bone age is at least one year behind chronological age. Also, height SDs are falling significantly, even compared to the standards for children with delayed puberty, and the velocity of body mass index is greater than zero.

We are just beginning to incorporate trend templates for intervening trends into $TrenD_x$. We plan to have the template for acquired GH deficiency triggered for three failed value constraints for the constitutional delay template: if the height SDs are falling too fast, if the weight SDs are falling too fast, or if the bone age is delayed more than 4 years behind chronological age. In this way $TrenD_x$ diagnoses a growth patient in the sequence: average growth, constitutional delay, growth hormone deficiency. This diagnosis sequence is familiar to expert pediatric endocrinologists.

Trends with Value Ranges Over Time

As noted in (EQ 1), value constraints specify that real-valued functions evaluated on the data within a certain time interval must stay between minimum and maximum bounds. This form was chosen to correspond to those constraints on laboratory values that make up much of the working knowledge of monitoring medical patient data.

For example, in the section on "Diagnostic Procedures in Liver Disease" in *Harrison's Principles of Internal Medicine* [Podolsky and Isselbacher 1991] describe laboratory results of patients with liver disorders. Of primary importance are the hepatic enzymes that aid decomposing and rebuilding of amino acids. One of these is aspartate transferase (AST). The authors note:

> in the patient with massive hepatic necrosis, there may be marked elevations [perhaps over 500 IU] in the early phase (i.e., 24 to 48 hours) , but by the time the patient is tested 3 to 5 days later the levels may be in the range of 200 to 350 IU [page 1309].

The levels of AST are ill-specified in both time and value primarily because the text aims to summarize a pattern for all hepatic necrosis patients.

We represent the above description as a trend template as shown below. Note that this is an intervening trend.

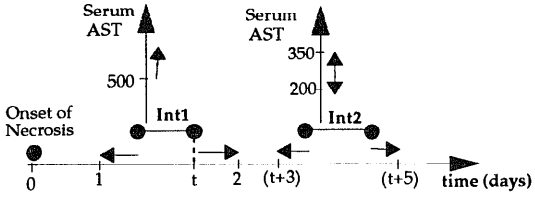

Figure 3 Trend template for serum AST pattern in hepatic necrosis. The time of the endpoint of Int2 is labeled t.

The lone landmark point for this template is the onset time of hepatic necrosis. Interval Int1 represents the period where AST levels are above 500 IU. Int1 begins 1 day or later after the onset of necrosis and ends (at time *t*) 2 days or sooner after the onset of necrosis. Interval Int2 represents the period where AST levels are between 200 and 350 IU. Int2 begins 3 days or later after *t* and Int2 ends 5 days or sooner after *t*.

Trends with Known Quantitative Models

Although value constraints were originally intended to represent constraints in highly uncertain areas like hepatic necrosis above, they may also be used in representing processes whose quantitative processes are better known. A value constraint can specify, for example, acceptable bounds for process data to match a time series model. Consider expressing the constraint that height standard deviations vary minimally from one point to the next as a first order autoregressive (AR(1)) model:

$$H_t = H_{t-1} + \varepsilon_t \qquad \text{(EQ 3)}$$

where H_t is a random-variable for the height SD at time t, and ε_t is a white noise random variable; $\varepsilon_t \sim N(0, \sigma)$. We can equivalently recast the AR(1) model of (EQ 3) as stating that the first difference of height SDs is a white noise variable:

$$H_t - H_{t-1} = \varepsilon_t \sim N(0, \sigma) \qquad \text{(EQ 4)}$$

When using this AR(1) model for trend detection of patient data, one must specify a confidence interval for believing that the actual patent data does conform to this model. For example, if we require a 95% confidence interval to establish a match, the value constraint is:

$$-1.96 \times \sigma \leq f(D) = (H_t - H_{t-1}) \leq 1.96 \times \sigma \qquad \text{(EQ 5)}$$

This is quite similar to the height SD value constraint on Int2 of Figure 1, with the exception that the value constraint in the figure divides f(D) by the time between the two data, since in general we can not assume that the height data are equally spaced as autoregressive models do.

One can similarly represent as a value constraint that a ceratin parameter P_t must be close to a curve model function g(t) over some time interval by using the value constraint:

$$-\delta \leq f(D) = (P_t - g(t)) \leq \delta \qquad \text{(EQ 6)}$$

The positive number δ is a noise threshold that may be chosen as in the previous equation. Examples of potentially useful g(t) are polynomials, exponentials, and trigonometric functions.

Clinical Trial with $TrenD_x$

Methods

We conducted a pilot clinical trial of $TrenD_x$ to evaluate its performance and to determine the weakness of the current representation and knowledge engineering. Data sets on 30 patients seen at the Division of Endocrinology at Children's Hospital were retrieved from the Clinician's Workstation (CWS), an on-line charting system [McCallie et. al. 1990]. The data sets included height, weight, sexual staging and bone age measurements. The patients were selected by filtering the problem list associated with each patient record in the CWS. Since this was an exploratory experiment rather than a rigorous test of efficacy, we specifically selected those problems and patient types for which we had engineered trend templates, as well as growth hormone deficiency, to explore how best to implement templates for intervening trends.

The first ten patient data sets were used as training sets. As errors in the performance of $TrenD_x$ on the training set were identified, we modified the trend templates by changing time ranges of the interval end points, by changing bounds of the value constraints, and by adding new intervals with new value constraints.

The remaining twenty data sets were used as test cases. These included patients with growth hormone deficiency, constitutional delay, average tempo of development, and early puberty. The data was read by $TrenD_x$ in chronological order. $TrenD_x$ recorded all the diagnoses and the age of the patient when they were considered or rejected. At the time of this trial, we had not developed trend templates for intervening trends like that for acquired growth hormone deficiency. All constitutional delay trend templates which $TrenD_x$ eventually rejected for any of these reasons:

1. the velocity of the height SD was too low,
2. the velocity of the weight SD was too low, or
3. the bone age was too far behind chronological age

were scored as diagnosing growth hormone deficiency when the constitutional delay trend template was ruled out.

Concurrently, a panel of three expert endocrinologists was given the same data sets and the task of diagnosing each patient. The endocrinologists were given the benefit of seeing the full data set at once rather than a point at a time. They too were required to judge the earliest age at which they could make their diagnosis. Note that this level of growth chart scrutiny is unusual in a busy general pediatric office. Several of the important contextual clues to the diagnosis that are usually gleaned from the patient history or laboratory results (e.g. results of a serum growth hormone test) were not available to either the clinicians or $TrenD_x$. Given the limitations of the data set, even in those cases where the panel consensus was different from the diagnosis stored in the CWS we took the panel consensus as the expert standard for a "correct" diagnosis.

Results

10 of the 20 patients were diagnosed by the panel as having one of these six disorders: normal growth, short stature, constitutional delay, early puberty, precocious puberty, and obesity. Of these $TrenD_x$ diagnosed 9 of 10 correctly. In 8 of the 9 correct diagnoses the clinicians reached the diagnoses at the same time as $TrenD_x$; in one case the clinicians were earlier. In the one case misdiagnosed by $TrenD_x$ the panel diagnosed constitutional delay, and the program diagnosed average prepubertal growth. This occurred because the patient's velocity of height standard deviation never crossed the lower bound of the value constraint on the average prepubertal growth trend template ($-\delta$ in Figure 1). We can correct this error by increasing the value of this lower bound.

The other 10 patients were diagnosed by the panel as having growth hormone deficiency. Of these, $TrenD_x$ diagnosed 5 of 10 correctly. In 3 of the 5 misdiagnosed cases a constraint on proportionality (such as body mass index) could have been used to correctly trigger growth hormone deficiency. For instance, $TrenD_x$ misdiagnosed one case of growth hormone deficiency as having constitutional delay. In this case, the clinicians noted that the weight and height did stay within a broad channel of their standard deviation. However they also noted that the weight standard deviation of the patient was creeping upwards at the same time that her height standard deviation was creeping downwards. As we had not encoded in the constitutional delay template any constraints on the proportionality of height and weight after infancy, $TrenD_x$ did not notice these subtle but significant opposing trends in height and weight. From this we have learned to add a constraint on proportionality to the constitutional delay template and to the acquired growth hormone deficiency template.

The results of this trial were encouraging in that they demonstrated that $TrenD_x$ could diagnose a few trends. However, the number of test cases, and the nature of the cases do not permit us to make any conclusions regarding the performance of $TrenD_x$ in general pediatric practice. We are planning larger trials in one of the primary care clinics at Children's Hospital to rigorously quantify the sensitivity and specificity of $TrenD_x$ as compared to clinicians.

Related Literature

We find that our pattern matching approach to trend detection fills a new niche in the work on monitoring

from time-ordered data. Several other projects may have results complementary to ours.

Traditional temporal logics [Allen 1984] have been used to encode the truth of logic propositions over time intervals. TrenD$_x$ extends this research by representing constraints on primary numerical data, as well as representing an entire process as phases.

Much of the work in diagnostic process monitoring has been in combined qualitative and quantitative simulations [Uckun and Dawant 1992]. This approach requires a domain where one can construct a causal model of the monitored process while TrenD$_x$ does not. Also, trend templates may supplement these qualitative simulation programs by indicating which sets of future qualitative states correspond to the same trend hypothesis. This may help to reduce branching of qualitative behaviors and thus improve monitoring efficiency.

Temporal abstraction programs [Shahar 1992] accept time-stamped laboratory data and create temporal intervals over which a parameter has attained a significant qualitative state (low, normal, markedly increasing, etc.). Unguided abstraction suffers in that it has very little context of what parameters are useful to abstract under a given hypothesis. For TrenD$_x$ this context is provided by the trend template of that hypothesis.

Conclusions and Future Work

A trend template represents a multi-variate trend in data from a monitored process (e.g. a medical patient) and incorporates both temporal and value uncertainty. A template may be anchored to specific dates on the calendar or to a specific patient age, or it may be offset from the onset time of an unexpected fault. Each interval of a trend template corresponds to a significant stage of the monitored process. Thus the constraints of a trend template may be more understandable to experts and knowledge engineers than differential equations or statistical curve-fitting models [Thissen and Bock 1990]. Among representations for disorders of monitored processes, the trend template is rare in requiring no pathophysiological model.

Our trend detection program TrenD$_x$ reached plausible diagnoses in most pediatric growth patients from a sample at Boston Children's Hospital. While these are promising results, we plan several epistemological improvements to make the program diagnose even more like an expert. Probabilistic bounds on value constraints would be useful for assigning numerical scores to the match of a datum to a template. Adding standard errors (due to measurement) on data values would make matching more flexible, and would allow more realistic projection of values over time [Dean and Kanazawa 1988].

We also plan improvements to the TrenD$_x$ diagnostic algorithms. The program should be able to ignore markedly aberrant data that do not fit a general trend. It should also be able to distinguish between competing hypotheses by ranking them. For two closely ranked disorders, TrenD$_x$ should request a laboratory test with high information content to distinguish between them, or suggest a waiting period after which the patient's data should differ under the two hypothesized disorders.

References

Allen, J. F. (1984). "Towards a General Theory of Action and Time." *Artificial Intelligence*, **23:**(2) 123-154.

Dean, T. and K. Kanazawa (1988). "Probabilistic Temporal Reasoning." *National Conference on Artificial Intelligence*, 524-528.

Haimowitz, I.J., and I. S. Kohane (1993). "Automated Trend Detection with Multiple Temporal Hypotheses." *International Joint Conference on Artificial Intelligence*, to appear.

Hamil, P. V. V., T. A. Drizd, C. L. Johnson, R. B. Reed, A. F. Roche and W. M. Moore (1979). "Physical Growth: National Center for Health Statistics Percentiles." *The American Journal of Clinical Nutrition*, **32**: 607-629

Kohane, I. S. (1987). *Temporal Reasoning in Medical Expert Systems*. MIT Laboratory for Computer Science technical report TR-389.

McCallie, D. P. Jr., D. M. Margulies, I. S. Kohane, R. Stalhut, and B. Bergeron (1990)."The Children's Hospital Workstation." *Symposium on Computer Applications in Medical Care*, 755-759.

Podolsky, D. K. and K. J. Isselbacher. (1991). "Diagnostic Procedures in Liver Disease." In *Harrison's Principles of Internal Medicine, Twelfth Edition.* McGraw Hill.

Shahar, Y., S. Tu and M. Musen (1992). "Knowledge Acquisition for Temporal Abstraction Mechanisms." *Knowledge Acquisition*, **1**:(4) 217-236.

Thissen, D. and R. D. Bock (1990). "Linear and Nonlinear Curve Fitting." *Statistical Methods in Longitudinal Research, Volume II: Time Series and Categorical Longitudinal Data.* Academic Press, Inc.

Uckun, S. and B.M. Dawant (1993). "Model-Based Reasoning in Intensive Care Monitoring, the YAQ Approach" *Artificial Intelligence in Medicine*, **5**:(1) 31-48.

Acknowledgments

Peter Szolovits, Mario Stefanelli, and Howard Shrobe have supplied valuable comments on this work. Doctors John Crigler, Samir Najjar, and Joseph Majzoub of Boston Children's Hospital Endocrinology Division kindly diagnosed the test cases. Khuram Faizan, Nadya Adjanee, and Phillip Le helped prepare the patient data for testing.