

A Clinical Trial of TrenDx: An Automated Trend-Detection Program

by

Phuc Van Le

July, 1997

@1997 Massachusetts Institute of Technology. All Rights Reserved.

A Clinical Trial of TrenDx: An Automated Trend-Detection Program

by

Phuc Van Le

Originally submitted to the Department of Electrical Engineering and Computer Science in September, 1996 in Partial Fulfillment of the Requirements for the Degree of Master of Engineering in Electrical Engineering and Computer Science

Abstract

We carried out a trial designed to assess the performance of a computer program called TrenDx at the task of diagnosing growth disorders in children, given a limited set of data. We compared the performance of TrenDx to that of physicians performing the same task. The task consisted of reviewing a growth chart of a patient and deciding whether the patient should be referred to a growth clinic for a possible growth problem, giving a preliminary diagnosis, and choosing the time which it would have been appropriate to refer the child. The test cases consisted of the height, weight, and bone-age data of 95 children that had been referred to the Boston Children's Hospital. The patient cases were organized into packets of 10 and distributed to physicians. Twenty two (22) physicians participated. Two gold-standards were used, the medical record diagnosis and the opinion of a pediatric endocrinologist. A scoring algorithm was devised based on the responses of the pediatric endocrinologists. The performance of the experts compared to the medical record indicated that it was possible to accomplish the task that we had designed. When compared to the medical record diagnosis, TrenDx and the physicians performed similarly in terms of referring patients, but the physicians chose the correct diagnosis more often. Compared to the experts, the physicians performed better than TrenDx in terms of referral decision and score.

Keywords: expert system, growth-disorder, evaluation, clinical trial, trend-template, medical decision-making, artificial intelligence, reasoning, diagnosis, process-monitoring, uncertainty reasoning.

This research has been supported by a fellowship from the Whitaker Foundation.

Acknowledgments

I would just like to thank everyone who has helped me to be what I am today and will become tomorrow.

- My parents and family are the most important people in the world to me. Without you, I would not exist. Thank you.
- My two thesis advisors, “Venerable Grandfather” Peter Szolovits and Isaac “Zak” Kohane, and Ira Joseph Haimowitz, who hired a naive sophomore who had no clue what he was getting himself into. I owe you three more than I can fit onto this page.
- I would also like to thank the people that participated in the trial, both the subjects and the experts.
- To Jane and Yao, who helped me see what being a doctor really meant.

- And to the two groups of people without whom my time at MIT would have been truly miserable :
 - To my friends on Conner 5 - May you always be able to just sit around and talk
 - To my volleyball buddies - 1,2,3 “Phil!”

- I would also like to thank the Whitaker Foundation for supporting me for the year and summer that I was working on this .

Table Of Contents

| | |
|--|------------------------------|
| 1. INTRODUCTION | ERROR! BOOKMARK NOT DEFINED. |
| 1.1. Pediatric Growth Monitoring | 1 |
| 1.2. Introduction to Trendx | 1 |
| 1.3. Evaluation | 1 |
| 1.4. Guide to this Thesis | 1 |
| 2. TRENDX | 1 |
| 2.1. TUP | 1 |
| 2.2. Temporal Constraints | 1 |
| 2.3. Value Constraints | 1 |
| 2.4. Trend Templates and Hypotheses | 1 |
| 2.5. Monitor Sets | 1 |
| 2.6. Previous Evaluations | 1 |
| 3. METHODS | 1 |
| 3.1. Packet Creation and Distribution | 1 |
| 3.1.1. Test Case Criteria | 1 |
| 3.1.2. Patient Record Collection | 1 |
| 3.1.3. Data-Entry | 1 |
| 3.1.4. Packet Creation, Distribution, and Return | 1 |
| 3.1.5. The Task | 1 |
| 3.2. Gold-Standard - Medical Record and Experts | 1 |

| | |
|---|-------------------------------------|
| 3.3. Performance Measures | 1 |
| 3.3.1. Comparison To Medical Record Diagnosis | 1 |
| 3.3.2. Comparison to Expert. | 1 |
| 3.3.3. Scoring Mechanism | 1 |
| 3.3.4. Other Comparisons | 1 |
| 3.4. Comparison to Other Evaluations | 1 |
| 4. TRENDX DEVELOPMENT | 1 |
| 4.1. Programming | 1 |
| 4.2. Modeling Growth | 1 |
| 4.2.1. Standards | 1 |
| 4.2.2. Description of Growth States | 1 |
| 4.2.3. Modeling | 1 |
| 4.3. Trend Template Refinement | 1 |
| 4.3.1. Establishing Performance Goals and Training Sets | 1 |
| 4.3.2. Error-Function Models | 1 |
| 4.3.3. Residual Mean Square Error vs. MAPE | 1 |
| 4.3.4. Thresholds for Triggering | 1 |
| 5. RESULTS | 1 |
| 5.1. Expert vs. Medical Record | 1 |
| 5.2. Comparison to Medical Record Diagnosis | 1 |
| 5.2.1. Decision-Breakdown by Abnormal Sub-Populations | Error! Bookmark not defined. |
| 5.2.2. Summary | 57 |
| 5.3. Comparison to Experts | 57 |
| 5.3.1. Referral Decision | 57 |
| 5.3.2. Preliminary Diagnosis | 59 |
| 5.3.3. Scores | 60 |
| 5.3.4. Summary | 62 |
| 5.3.5. Multiple Comparisons of Consensus Cases | 62 |
| 5.3.6. Summary | 63 |

| | |
|--|------------|
| 5.4. Variations in Threshold Triggering | 64 |
| 5.4.1. Results of Raising Threshold Triggering Values | 64 |
| 5.4.2. Results of Lowering Threshold Triggering Values | 65 |
| 5.4.3. Ignoring Certain Error-Scores | 65 |
| 5.5. Referral Decision Timing | 66 |
| | |
| 6. DISCUSSION | 68 |
| | |
| 6.1. Analysis of Results | 68 |
| 6.1.1. Expert vs. Medical Record Gold-Standard | 68 |
| 6.1.2. Performance of Test Subjects vs. Medical Record Diagnosis | 69 |
| 6.1.3. Performance of Test Subjects vs. Expert Gold-Standard | 69 |
| 6.1.4. Referral Timing | 70 |
| | |
| 6.2. Performance of Trendx | 70 |
| 6.2.1. Worse than Physicians (?) | 70 |
| 6.2.2. Reasons for Poor Performance | 71 |
| | |
| 6.3. Limitations of This Trial | 74 |
| | |
| 7. CONCLUSION | 77 |
| | |
| 7.1. Summary of Results | 77 |
| | |
| 7.2. Lessons Learned | 77 |
| | |
| 7.3. Future Work | 78 |
| | |
| 8. APPENDIX A - PATIENT AND SUBJECT RESULT TABLES | 79 |
| | |
| 9. APPENDIX B - PACKET DIRECTIONS / SAMPLES | 125 |
| | |
| 10. APPENDIX C - TREND TEMPLATE LISP CODE | 128 |
| | |
| 11. REFERENCES | 127 |

Table of Figures and Tables

| | |
|---|-------------------------------------|
| FIGURE 1: TWO LANDMARK POINTS IN THE LIFE OF AN INDIVIDUAL | 1 |
| FIGURE 2: AN INTERVAL REPRESENTING THE LIFE OF AN INDIVIDUAL. | 1 |
| FIGURE 3: A NOW-BASED TREND-TEMPLATE INTERVAL. | 1 |
| FIGURE 4: EXAMPLES OF POSSIBLE VALUE CONSTRAINTS | 1 |
| FIGURE 5: TREND-TEMPLATE FOR NORMAL MALE GROWTH | 1 |
| FIGURE 6: SAMPLE HUMAN SUBJECT RESPONSE | 1 |
| FIGURE 7: SAMPLE GOLD-STANDARD RESPONSE | 1 |
| FIGURE 8: CURRENT TREND-TEMPLATE FOR NORMAL MALE GROWTH | 1 |
| FIGURE 9: PREVIOUS TREND-TEMPLATE FOR NORMAL MALE GROWTH(HAIMOWITZ) | 1 |
| FIGURE 10: COMPARISON OF SIMPLE AND COMPOUND VALUE CONSTRAINTS TO CONSTRAINT-BASED TRENDX | 1 |
| FIGURE 11: POSSIBLE COMPOSITE ERROR-FUNCTIONS | 1 |
| | |
| TABLE 1: TRENDX MATCHING RESULTS ON TERTIARY CARE PATIENTS, FROM (HAIMOWITZ) | ERROR! BOOKMARK NOT DEFINED. |
| TABLE 2: MEDICAL RECORD DIAGNOSES OF TRIAL CASES | 1 |
| TABLE 3: TABLE OF RESULTS FOR A SAMPLE PATIENT | 1 |
| TABLE 1: MEDICAL RECORD DIAGNOSES OF TRIAL CASES | 1 |
| TABLE 2: EXPERT DECISION TO REFER VS. MEDICAL RECORD DIAGNOSIS | 1 |
| TABLE 3: DISORDER POPULATION REFERRAL AND DIAGNOSIS, EXPERT VS. MEDICAL RECORD | 1 |
| TABLE 4: TRENDX DECISION VS. MEDICAL RECORD DIAGNOSIS | 1 |
| TABLE 5: ALL PHYSICIANS VS. MEDICAL RECORD DIAGNOSIS | 1 |
| TABLE 6: PRE-RESIDENCY SUBJECTS VS. MEDICAL RECORD DIAGNOSIS | 1 |
| TABLE 7: POST-RESIDENCY SUBJECTS VS. MEDICAL RECORD DIAGNOSIS | 1 |
| TABLE 8: DISORDER POPULATION REFERRAL AND DIAGNOSIS RESULTS, TRENDX VS. MEDICAL RECORD | 1 |
| TABLE 9: DISORDER POPULATION REFERRAL AND DIAGNOSIS RESULTS, | 1 |
| TABLE 10: DISORDER POPULATION REFERRAL AND DIAGNOSIS RESULTS, | 1 |
| TABLE 11: DISORDER POPULATION REFERRAL AND DIAGNOSIS RESULTS, POST- RESIDENCY VS. MEDICAL RECORD | 1 |

| | |
|--|---|
| TABLE 12: TRENDX DECISION VS. EXPERT GOLD-STANDARD | 1 |
| TABLE 13: TRENDX DECISION VS. EXPERT AND MEDICAL RECORD CONSENSUS | 1 |
| TABLE 14: PHYSICIANS VS. EXPERT GOLD-STANDARD | 1 |
| TABLE 15: PHYSICIANS VS. EXPERT AND MEDICAL RECORD CONSENSUS | 1 |
| TABLE 16: TEST SUBJECT PRELIMINARY DIAGNOSIS MATCHES TO EXPERT DIAGNOSIS | 1 |
| TABLE 17: TEST SUBJECT SCORES | 1 |
| TABLE 18: AVERAGE SCORE BY DISORDER SUB-POPULATION | 1 |
| TABLE 19: CONSENSUS AND SINGULAR REFERRAL DECISIONS | 1 |
| TABLE 20: RESULTS OF RAISING TRIGGERING THRESHOLDS | 1 |
| TABLE 21: RESULTS OF LOWERING TRIGGERING THRESHOLDS | 1 |
| TABLE 22: RESULTS OF IGNORING FIRST NON-TRIVIAL POINT | 1 |
| TABLE 23: RESULTS OF IGNORING INFANT SCORES | 1 |
| TABLE 24: TIMING OF REFERRALS FOR TRENDX VS. EXPERT | 1 |
| TABLE 25: TIMING OF REFERRALS FOR HUMAN SUBJECTS VS. EXPERT | 1 |

1. Introduction

In many domains, experts can judge the state of a process by examining the data produced by the process and then matching these data to stereotypical patterns specific to different states. Haimowitz defines the term **trend** as a clinically significant pattern in a sequence of time-ordered data (Haimowitz). Thus, **trend-detection** is the task of judging the state of a process by matching the data produced by the process to different trends. This trend-detection is applicable in many domains. We are particularly interested in evaluating its application to the domain of medicine, where the task of diagnosis can be viewed as matching a patient's findings to trends that are typical of certain conditions.

The computer program *TrenDx* was developed by Haimowitz based on the premise - "that a computer program with knowledge of time-varying constraints on measured data can be used for automated trend detection." (Section 1 Haimowitz). *TrenDx* was tested in a limited manner as part of its original development. This research is a more stringent evaluation of *TrenDx* at the task of diagnosing growth disorders in children from their height, weight, and bone-age data.

1.1. *Pediatric Growth Monitoring*

Health care systems have been under intense pressure to become more efficient and cost-effective. This has had many consequences, including forcing physicians to see more patients and spend less time with each individual patient. Often, these time pressures have led to care that is less than optimal. The direct experience of one of the advisors of this thesis has found that children with growth problems are not being diagnosed, and therefore treated, in a timely fashion. Similarly, patients with normal growth are sometimes referred to tertiary care centers for expensive work-ups because their physicians misdiagnose normal patterns of growth.

Being able to diagnose referrals correctly more often would improve health care because patients with abnormal growth would be diagnosed and treated before their conditions became grossly apparent and possibly untreatable. Reducing mortality and morbidity while lowering the cost of care by eliminating needless referrals are the ultimate goals of this research. However, these goals must be accomplished without increasing the physician's workload.

Thus, the task of pediatric growth monitoring is an important one in which the performance of pediatricians can possibly be improved upon by applying automated trend-

detection. In the growth clinic at the Boston Children's Hospital, pediatric endocrinologists routinely quiz one another by asking their peers to make a diagnosis solely from the information contained on a patient's growth chart. Furthermore, Becker notes that "Thus, the monitoring of linear growth is a remarkably cost-effective screening for the documentation of good health or for determining the presence or severity of chronic disease." (Becker)

1.2. Introduction to TrendX

TrendX is described in more detail in section 2. It is also described in its entirety in (Haimowitz; Haimowitz and Kohane; Haimowitz and Kohane). Briefly, knowledge-engineers and domain specialists outline stereotypical patterns in temporal data using the modeling language incorporated into TrendX. These patterns, called trend-templates, consist of partially ordered temporal intervals, each with constraints on all the data that fall into that particular time interval. Data within each interval are matched to the constraints associated with that interval using linear regression techniques, producing an error-score which indicates how well the data match to the particular trend-template. Trend-templates are grouped into competing sets called monitor sets, from which the best-scoring trend-template is considered to be the current hypothesis or diagnosis.

1.3. Evaluation

A rigorous evaluation is an essential part of the development of any system (Heathfiled and Wyatt; Waterman). A statement of the exact goals of this evaluation is necessary.

- We wish to assess the performance of a computer program, TrendX, at the task of recognizing growth disorders in children from a limited data set and then referring the child to a specialist if appropriate.

The immediate goal of the evaluation is to show that TrendX can perform this task with some expertise. The long-term goal of the development of TrendX is to create a smart monitoring system that can detect the state of a process by recognizing significant trends in the data produced by the process and cause some type of action to be taken if the data suggest an undesirable state. In addition, we wish to improve the state of expert-systems concerning the incorporation of temporal knowledge into their knowledge base.

Evaluations of decision-support systems can be divided into two categories, laboratory trials and field trials. Laboratory trials are carried out during the earlier phase of

development of a system and are characterized by more controlled conditions such as retrospectively chosen cases, “clean” data, and users who are very familiar with the systems. A handful of systems are carried through to the stage where field trials are appropriate. In field trials, the environment is much less controlled, causing new problems to arise. These generally include a wider range of cases, novice users, a larger setting, different outcome measures, and potential legal and ethical considerations concerning the use of the output generated by these programs.

At this stage in the development of Trendx, a laboratory trial of the performance of the program is appropriate. Thus, we have designed a retrospective clinical trial of Trendx in which the performance of Trendx is compared to human experts - physicians. Both Trendx and the human subjects, collectively referred to as the test subjects, are given the complete set of height, weight, and bone-age data available for a patient. The test subjects must decide if they would recommend that the patient be referred to the endocrine division to be worked-up for a possible growth problem. The test subjects are also asked to give a preliminary diagnosis and choose the age at which the referral should have been made. All answers are measured against 2 gold-standards - the diagnosis written in the patient’s medical record and the opinion of a pediatric endocrinologist.

This task can be viewed as that of one physician giving a second opinion to a colleague who suspects that one of his or her patients has a growth problem. The task is also analogous to that of the individual in a managed care organization who has to decide whether a referral to a tertiary care center is warranted. In section 3.4, the characteristics of this trial are compared to some of the other types of evaluations of decision-support systems that have been carried out.

1.4. Guide to this Thesis

Section 2 describes Trendx and the trend-representation language which Trendx uses. Intervals, value constraints, trend-templates, and monitor-sets are all explained well enough for someone unfamiliar with the program to understand the work presented in this thesis. The section also discusses the most recent evaluations of Trendx and some of their weaknesses.

In section 3, all the work not directly related to the development of Trendx is presented. This work includes the collection of test cases, the transcription of the data into electronic format, and the creation of test packets for distribution to the participants. The section also describes the task which the participants in the trial are asked complete. The

two gold-standards for this evaluation are described, as well various measures which were used to evaluate the performance of TrenDx and the 22 human participants. Finally, a comparison to other evaluations of expert-systems / decision-support systems is made.

The development of TrenDx comprises the bulk of Section 4. This covers the programming improvements into TrenDx, creation of trend-templates that model the different growth states/disorders involved in the trial, and the engineering of the trend-template parameters to achieve a desired level of performance.

Section 5 presents the results of the trial. Some of these include the results of comparisons between the gold-standards and the decisions made by the subjects, comparisons between the gold-standards themselves, and the results of various changes to the mechanism used to trigger a referral.

The discussion of the results can be found in section 6. Conclusions about the trial and the future work / uses of TrenDx are in section 7.

Appendix A - Patient and Subject Result Tables, contains the entire listing of the results of the trial, by patient case and by human subject. Appendix B - Packet Directions / Samples , includes the directions presented to the participants and a sample chart similar to the ones on which they indicated their responses. Appendix C - Trend Template LISP Code, shows the LISP code for all of the trend-templates used the the trial, for those who are interested in such things.

There are several conventions used in this thesis. For example, the word subjects is used to refer to both TrenDx and the physicians that participated in the trial. It does not include the pediatric endocrinologists who provided one of the gold-standards. The terms human subjects, participants, volunteers, and physicians all refer to the human subjects who participated in the trial, even though not all of them are physicians. The gold-standard provided by the pediatric endocrinologists is also referred to as the expert opinion or the expert decision. Patient cases, cases, and patients all refer to the patient cases that were reviewed by the subjects and the experts in this trial. Finally, references to tables and figures will usually only contain the caption name and number, such as Table 1. However, if, there is a possibility for ambiguity, then the title of the reference will also be included, such as Table 1:TrenDx matching results on tertiary care patients, from (Haimowitz).

2. TrendDx

The most detailed description of TrendDx can be found in (Haimowitz). The brief description presented here is provided to supplement the discussion of improvements to the program and the engineering choices made in the representation of growth disorders that is presented later in this thesis.

TrendDx diagnoses trends by matching time-ordered process data to the competing trend templates in each monitor set assigned to that process. TrendDx begins matching by instantiating each trend template for the monitored process. TrendDx then computes all temporal worlds in which the currently interpreted data may be assigned to intervals of the trend template. Each temporal world represents a different hypothesis for the same trend template. For each hypothesis, TrendDx assigns the data to the appropriate trend template intervals and computes the matching scores of the relevant value constraints. The value constraint scores are combined to an overall error score for each hypothesis. Finally, the top hypotheses for each trend template are maintained via a beam search. The output of TrendDx is a list of the top hypotheses for each trend template within a monitor set, with the score of each hypothesis. (Section 4 Haimowitz).

2.1. TUP

TrendDx manipulates temporal assertions and queries using the Temporal Utility Package, or TUP, developed by Kohane (Kohane). TUP is a set of temporal utilities which allow TrendDx to represent time points and intervals, as well as reason about uncertainty in temporal distances. For example, TUP allows the expression of time intervals with uncertain endpoints. Furthermore, TUP provides the ability to deal with alternate temporal worlds. Alternate temporal worlds are contexts in which different temporal assertions apply. For example, say a user specified that Event A occurred sometime between January 1, 1990 and December 31, 1990. Then say that another event, Event B, occurred on July 1, 1990. From the known information, Event B could occur before, at the same time as, or after Event A. TUP allows the formulation of alternate temporal worlds wherein each of the relationships between Event A and Event B is asserted.

The ability to deal with time is an important aspect of any medical decision-support system. This is especially true in the domain of pediatric growth monitoring. One of the conclusions of the INTERNIST-1 program was that the inability to incorporate temporal

information into the program was one of its major weaknesses (Miller, Pople and Myers). The use of TUP enables TrendX to incorporate this type of knowledge into its models and reasoning.

2.2. Temporal Constraints

The temporal aspects of a trend-template include **landmark points** and **intervals**. Landmark points represent significant events during a process. For example, BIRTH and DEATH would be considered landmark points in the process of a person's life. The temporal distance between landmark points can be specified with a set of lower and upper bounds (MIN MAX), indicating the minimum and maximum difference in time between the two points. For example, for a person who lived exactly 80 years, the temporal distance between their BIRTH and DEATH would be represented by ((years 80) (years 80)). This specifies that both the minimum and maximum distance between the BIRTH and DEATH landmark points is eighty years, meaning that exactly eighty years separated the two points. To represent the fact that a particular person died sometime between the ages of 13 and 20, the (MIN MAX) set would look like ((years 13) (years 20)). Figure 1 shows a timeline with the landmark point DEATH occurring 13 to 20 years after BIRTH.

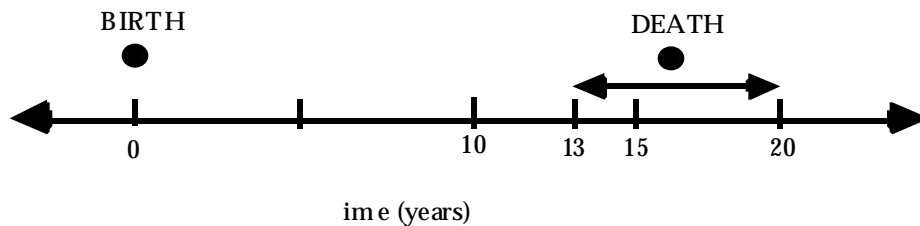


Figure 1: Two landmark points in the life of an individual

Intervals are used to represent different phases of a process. In TrendX, intervals are represented by a Begin point and an End point. The temporal distance between the two represents the duration of the interval. Begin and End points can also have uncertainty ranges associated with their relation to another time point, either a landmark point or an interval Begin/End point. In Figure 2 we extend the previous example by adding an interval representing the time period over which the person lived.

As suggested before, interval Begin/End points can be defined relative to other interval Begin/End points. One common relationship between two intervals occurs when

one interval directly follows another. In that case, the End point of the first interval is equal in time to the Begin point of the second.

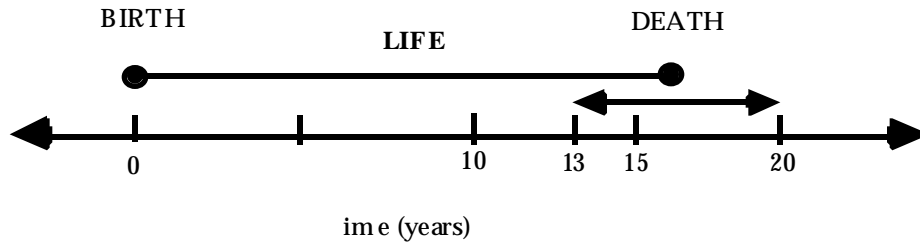


Figure 2: An Interval Representing the Life of an Individual.

In some cases, it is more relevant to view time backwards beginning from a particular time point. Most often, the particular time point is the present. For example, to determine whether a child is obese, the state of the child at the present time is much more important than the child's state 2 years ago. Similarly, to determine whether the child has a fever, it is the current temperature of the child that is relevant. The need to model this view of time led to the development of what is called a Now-Based trend-template.

A Now-Based trend-template has an additional anchor point called 'now' which represents the most recent data point. The 'now' point is updated to be equal in time to each new data point that is processed. This allows a user to design a trend-template and set the interval Begin/End points relative to 'now.' Figure 3 is an example of a Now-Based trend-template that models a child with fever. The duration of the Fever interval is specified relative to 'now' and extending back somewhere between 30 minutes to 2 hours.

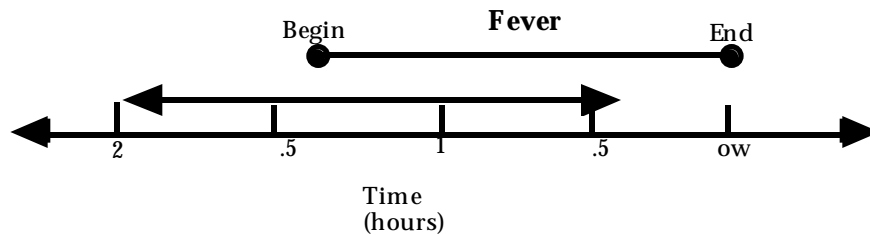


Figure 3: A Now-Based Trend-Template Interval.

2.3. Value Constraints

A value constraint is composed of two main components. The first component is a function that maps the data in the interval to a time-indexed real-valued sequence of numbers. The second component is a linear regression model describing the pattern of the output of the first component.

The first component can be as simple as a function that simply returns the numerical value of each time-stamped datum. Or it may return some more complicated function of several data points of different types. An example of a simple function may just return a sequence of the height Z-scores, which represent how many standard deviations the child is from the average height at that particular age. A more complicated function is one which returns the ratio of weight to weight for height-age, which we call Build.

The second component, which we call the error function, can be one of a set of up to 2nd order polynomial functions. The constant and 1st order polynomial functions can even specify the value and slope of the function to be matched against the data. In other words, the second component can specify that the sequence returned by the first component should be matched to a constant of known or unknown value, a line of some known or unknown slope, or a 2nd order polynomial curve with first and second derivatives positive or negative. Figure 4 gives some examples of value constraints.

Examples of Possible Value Constraints


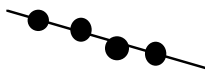
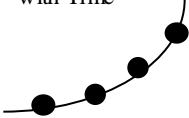
| Concept | Component 1 | Component 2 |
|--|----------------|---------------------------|
|  <p>Pulse Constant at 60 bpm</p> | Pulse | (Constant 60) |
|  <p>Blood Pressure Linear and Falling</p> | Blood Pressure | (Linear (D1 -)) |
|  <p>CPU speed Quadratic with Time</p> | CPU Speed | (Quadratic (D1 +) (D2 +)) |

Figure 4: Examples of Possible Value Constraints

The sequence is matched to the error function, producing the residual mean square error. This is repeated for all the value constraints of all intervals of the trend-template. The residuals are combined by using a weighted average of the fits to each value constraint, where the weight is proportional to the number of data points in each value constraint.

Haimowitz suggests using the Mean Absolute Percent Error, or MAPE, because parameters with larger ranges will have a larger variance of residuals (Section 4.4.1 Haimowitz). Using the percent error allows one to combine the errors from different value

constraints correctly. However, the use of MAPE is not applicable when the expected value of the parameter is zero. This is discussed in section 4.3.

Originally, TrendX used simple upper and lower bounds to determine whether a data point matched to an interval well (Haimowitz and Kohane; Haimowitz and Kohane). Only one trend-template was active at any single time and it was considered the current hypothesis. When a data point exceeded the upper or lower thresholds, another trend-template was triggered or some other action, such as an alarm, was taken.

This style of value-constraint matching was known as Constraint-Based TrendX. It was originally designed to mimic the stream of thought of an expert. For example, an expert would start off with the hypothesis that the child was normal. Then, if the child's height was too low, the expert would then discard the current hypothesis and consider Constitutional Delay of Puberty as the current hypothesis.

Constraint-Based TrendX suffered from many drawbacks. Similar to other threshold-trigger systems, Constraint-Based TrendX was brittle. For example, if the lower threshold on a value was -2.0, then a value of -2.1 would cause the current hypothesis to be discarded, while a value of -2.0 would not. In addition, there was no difference between having a data point that was exactly normal and one which fell just within the allowable threshold. Regression-Based TrendX was developed to remedy some of those problems.

2.4. Trend Templates and Hypotheses

A trend-templates represents an overall state of a process. It is comprised of a partially ordered set of intervals, each with one or more value constraints associated with them. When a trend-template is instantiated for a patient, a temporal context is created in which the BIRTH landmark point of the patient is anchored to the time and date that the patient was born. When alternate temporal worlds are possible, the context branches, producing multiple child contexts that represent each of the possible temporal worlds.

Thus, each hypothesis for a patient consists of a trend-template, a temporal context, and an assignment of the patient data to the intervals of the trend-template that is dependent on the temporal context.

Figure 5, is an example of the complete trend-template for Normal growth in males. Notice that this trend-template only tries to model life up to and a little beyond the point where growth stops.

Trend Template for Normal Male Growth

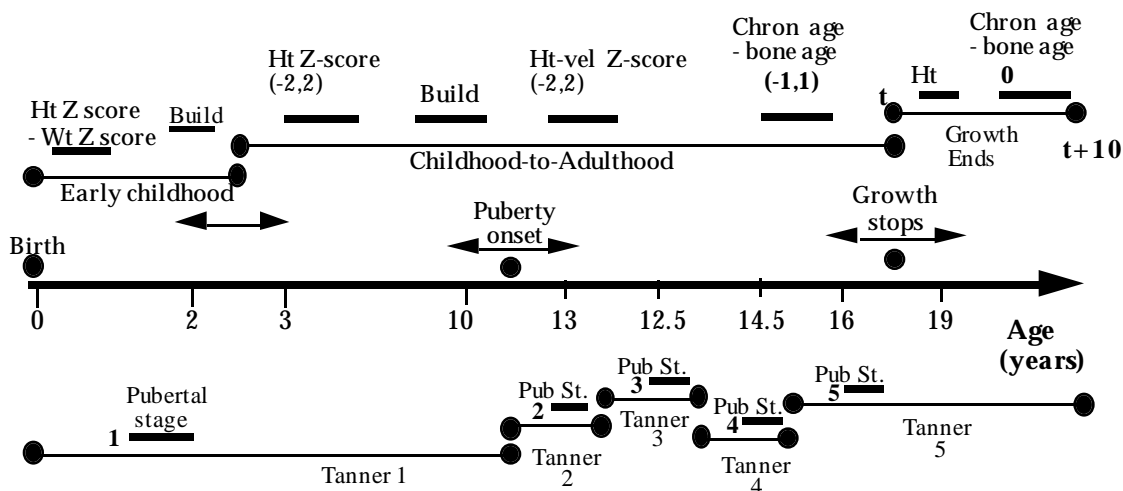


Figure 5: Trend-Template for Normal Male Growth

2.5. Monitor Sets

Trend-templates are grouped into competing sets called monitor sets. Haimowitz defines a monitor set as a set of trend templates forming a clinical context for monitoring. Error scores from each trend-template can be compared with the error scores from the other trend-templates within the monitor set to determine which trend-template has the best fit to the data.

Since a monitor set describes a group of competing trend-templates, it typically consists of a ‘normal’ trend-template that represents the process in its normal state and one or more ‘abnormal’ trend-templates that represent fault states.

2.6. Previous Evaluations

In (Haimowitz and Kohane), the performance of constraint-based TrendDx was evaluated, using a panel of three pediatric endocrinologists as a gold-standard. Out of the 20 test cases, 14 were diagnosed correctly by TrendDx.

Haimowitz later performed an evaluation of the regression-based TrendDx as part of his thesis (Section 5.1 Haimowitz). This version of TrendDx will be referred to as thesis-TrendDx. In the evaluation of thesis-TrendDx, two sets of test cases were used. The first consisted of 30 cases. Of these, there were 4 Normal, 10 had Constitutional Delay of Puberty, 3 were diagnosed with Early Puberty, and 13 suffered from Growth-Hormone

Deficiency. The monitor set consisted of trend-templates for Normal growth, Early Puberty, and Constitutional Delay of Puberty. For the Growth-Hormone deficiency cases, a diagnosis of Constitutional Delay of Puberty was considered correct.

Table 1 is adapted from (Haimowitz). It shows the performance of thesis-TrenDx on the 30 cases. The column labeled ‘Persistent Gap,’ ‘Single Gap,’ and ‘Union’ describe the criteria used to determine whether the Constitutional Delay or Early Puberty trend-templates overtake the Normal Growth trend-template. In essence, ‘Persistent Gap’ requires that the Normal Growth trend-template scores somewhat worse than either of the other two trend-templates for two consecutive time points. ‘Single Gap’ requires that the Normal Growth trend-template scores *significantly* worse than either of the other two trend-templates. ‘Union’ is the union of the both Persistent and Single gap triggering mechanisms. The use of Persistent Gap is to try to reduce some of the brittleness inherent in using threshold-triggering, as discussed in Section 2.3. Formally:

- Persistent gap: For two consecutive visits, the best hypothesis for P scored 0.8 or less times the best hypothesis for Normal Growth.
- Single gap: For one visit, the best hypothesis for P scored 0.6 or less times the best hypothesis for Normal Growth.
- Union: Either Persistent Gap or Single Gap.

Where P is the trend-template for either Constitutional Delay of Puberty or Early Puberty. The sensitivity and specificity of thesis-TrenDx was calculated for the diagnosis of both Constitutional Delay of Puberty and Early Puberty as follows:

$$\text{sensitivity for P} = \frac{\text{(number with P and TrenDx triggers P)}}{\text{(Number with P)}}$$

$$\text{specificity for P} = \frac{\text{(number without P and TrenDx does not trigger P)}}{\text{(Number without P)}}$$

Notice that the sensitivities are low, ranging from 0.0 to 0.66, while the specificities are good, ranging from 0.77 to 1.0. Clearly, the results could have been changed by changing the triggering criteria. Lowering the threshold value used to trigger the alternate trend-templates would most likely have increased the sensitivity at the cost of decreasing specificity. This is a common tradeoff that is described by a Receiver Operator Characteristic Curve(Pagano and Gauvreau).

Table of Results from the Previous Evaluation of TrenDx

| Disorder | # | Persistent Gap | | | Single Gap | | | Union | | |
|---------------------|----|----------------|---------------|--------------|------------|---------------|--------------|-------|---------------|--------------|
| | | Norm | Cons Delay | Early Pub | Norm | Cons Delay | Early Pub | Norm | Cons Delay | Early Pub |
| Normal | 4 | 4 | 0 | 0 | 3 | 0 | 1 | 3 | 0 | 1 |
| Cons. Delay | 10 | 7 | 3 | 0 | 5 | 5 | 1 | 4 | 6 | 1 |
| Early Puberty | 3 | 3 | 0 | 0 | 1 | 0 | 2 | 1 | 0 | 2 |
| GH Deficiency | 13 | 4 | 6 | 4 | 5 | 6 | 3 | 4 | 6 | 4 |
| Cum Sensitivity | 30 | | 0.39 | 0.00 | | 0.48 | 0.66 | | 0.52 | 0.66 |
| Cum. Specificity | 30 | | 1.00 | 0.85 | | 1.00 | 0.81 | | 1.00 | 0.77 |

Table 1: TrenDx matching results on tertiary care patients, from (Haimowitz)

The second set of test cases used consisted of 20 cases taken from the files of a general pediatrician. These cases were considered ‘normal’ by the pediatrician, but were not reviewed by any pediatric endocrinologists. The table presenting the results of those 20 additional cases (Table 2 Section 5.1 Haimowitz) has several small inconsistencies and will not be included here. The results of the second set of test cases has very little effect on the Cumulative Specificity of the program.

There are several characteristics of the trial that incorporate bias or weaken the ability to draw conclusions from the results. One of the most obvious problems with the trial is the small number of trend-templates. There were only three trend-templates used - Normal Growth, Early Puberty, and Constitutional Delay of Growth. The program was considered correct if it chose Constitutional Delay when the patient had Growth-Hormone Deficiency. In truth, combining the diagnosis of Growth-Hormone Deficiency with Constitutional Delay of Puberty is very suspect. While both conditions exhibit short stature as a result of delayed growth, Constitutional Delay of growth is generally considered a benign condition and not worthy of a growth clinic referral unless the patient has a very extreme case. On the other hand, Growth Hormone Deficiency is a true, secondary disturbance of growth that can be related to even more serious problems such as a

craniopharyngioma (brain tumor). A second problem with the trial is that some of the Growth-Hormone Deficiency cases had been used previously as test cases in a previous trial (Haimowitz and Kohane) of Constraint-Based TrenDx. Thus, the test was biased because those cases had influenced the previous development of the Regression-Based TrenDx. However, the design of the test was appropriate to the level of development of the program. Formal evaluations, such as double-blind, comparative studies are less appropriate at early levels of development because the experts should probe the inference engine and knowledge base of the program, not just be worried about final program results.

In summary, the results of the evaluation were promising. However, they indicated that further work would be necessary to improve the performance of the program and that a more complete evaluation of the program would be appropriate at that time.

3. Methods

We have conducted an experimental trial of a revised version of TrenDx using 95 newly-collected patient cases taken sequentially from the patients referred to the Endocrine Division at the Boston Children's Hospital. The cases were screened for inclusion into the trial, then the data in the cases were transcribed and placed into packets for each human subject. Each packet contained growth charts for 10 test cases, distributed to try to achieve an even distribution of test cases. The packets were distributed as participants were found. The participants consisted of physicians, medical students, and a registered nurse. Over 80 packets were created and distributed, but only 22 were returned. The medical record diagnosis for each of the cases was obtained and used as one gold-standard. A pediatric endocrinologist not involved with the development of the program provided a second gold-standard for the trial.

TrenDx was updated and improved, independent of the test cases that were going to be used for the trial. New trend-templates were designed to try to take into account a new variation of triggering as well as improve over the performance of thesis-TrenDx. In fact, this trial is the first formal test of any trend-templates other than those for Normal Growth, Early Puberty, or Constitutional Delay of Puberty. Several problems with the use of 'now-based trend-templates were uncovered and solved. The development of TrenDx is discussed in Section 4.

3.1. Packet Creation and Distribution

3.1.1. Test Case Criteria

The occurrence of growth abnormalities in the general population is too low to use it as a test population - for example, Congenital Growth Hormone Deficiency occurs in approximately 1 out of 16,000 people (Kaplan). However, arbitrarily picking a certain number of cases of different pathologies is difficult to justify because the numbers would not reflect the relative frequency of the different pathologies that are referred to the growth clinic. Consequently, we decided to take cases randomly chosen from patients that had been referred to the Division of Endocrinology at the Boston Children's Hospital. This population has both a high proportion of abnormal patients, as well as normal patients that had some characteristics of abnormal patients. Some of the cases were not referred to the growth clinic because of a suspicion of growth abnormality, but were referred for some other reason. We consider these cases to be "normal" as well.

Recall that the motivation of the program was to improve the performance of physicians in the domain of pediatric growth monitoring by helping to diagnose children with growth disorders and by reassuring the physician that a normal child is truly normal. One might argue that using a referral population only tests the ability of the test subjects to do the latter task and not the former - i.e. that using this test population only allows us to catch patients with normal growth who were referred incorrectly and that it does not allow us to catch patients with abnormal growth who were not referred. However, a child who does have a true growth disorder will become more symptomatic as time progresses. In fact, most of the disorders cause the children to fall 2 or 3 standard deviations below the mean height for that age and to keep falling away from their peers. Thus, it is a somewhat simple task to recognize a child with a growth disorder if the child has been suffering from it long enough. Our aim is to improve the timing of the diagnosis and referral to minimize morbidity in these children.

Other evaluations have used referral cases in a similar fashion. Heckerman uses referral cases in his evaluation of the Pathfinder program (Heckerman and Nathwani). In an evaluation of four decision-support systems by (Berner et al.), the test cases consisted of referrals to a "...group of 10 nationally recognized consultants in the fields of general internal medicine, eight subspecialties of internal medicine, and neurology..." They chose to use the referral cases to ensure that the cases were diagnostically challenging.

3.1.2. Patient Record Collection

To decide whether to accept a patient case into the trial, the physical record was scanned and the age at which the child was first referred to the clinic was noted. If the record contained data for at least three time points before the referral date, we tentatively accepted the case. A second criterion was that only patients that were referred more than one year ago were accepted. This was done to allow the true clinical outcome of the patient to be used as one of the standards. We then screened out previously diagnosed cancer patients. There were two reasons for this decision. First, both cancer and its treatments have complex effects on growth. Second, patients receiving cancer treatment were assumed to be under close clinical observation and the original motivation of the trial was to catch cases which were diagnosed late because of time pressures on the pediatrician. Of the patient cases that were screened, approximately 70% had enough data and the right background for us to accept the case. We collected approximately 120 patient records and numbered them consecutively starting at 2000.

3.1.3. Data-Entry

All height, weight, and bone-age data of the child were entered into a spreadsheet. Recall that only the data available before the date the child was seen at the clinic were used in this trial. The growth clinic usually acquires the information from the referring doctor by calling the referring doctor's office and verbally transcribing the data or by receiving a faxed copy of the patient's growth chart. Therefore we took photocopies of the patient's growth charts and the verbal transcriptions.

Once the information was entered into a spreadsheet, it was transformed into TrenDx-readable LISP code by a series of programs. Here is some example code:

```
(make-patient 'BOY-PATIENT :id 3 :dob "1/1/80"  
  :name "Fake patient ID# 3")  
(add-patient-datum 'height 3 84 :age 2)  
(add-patient-datum 'weight 3 12.5 :age 2)  
(add-patient-datum 'height 3 93 :age 3)  
(add-patient-datum 'weight 3 14.5 :age 3)
```

The above code creates a male-patient whose date of birth is January 1, 1980 and assigns the patient the id number 3. It then adds 4 data elements to the patient - 2 height and 2 weight data, taken at ages 2 and 3. Data that have the same time-stamp, such as the height and weight pair taken at age 2, are considered a data-cluster and are processed together.

At this stage, approximately 8 of the 120 cases had to be removed from the trial because some portion of the record was unreadable. From the remaining 112 cases, the first 100 were chosen to be included in the trial and distributed. The others were not used.

3.1.4. Packet Creation, Distribution, and Return

Each patient's data were displayed on a growth chart (**Appendix B - Packet Directions** / Samples). The charting process was automated by writing a Hypercard application that automatically plotted the data on either an infant chart (age 0-3) or a childhood chart (age 2-18), or both if appropriate. The test cases were then distributed among the packets in a way to try to equalize the number of test subjects that saw each case, while preventing any two packets from containing the same 10 individual patient cases. However, since packets were not returned frequently, the number of responses per patient case varies significantly.

The human subjects were recruited in several different manners. Many of the subjects were physicians at the Boston Children's Hospital. They were asked to participate and those who agreed were handed packets with return envelopes. Other subjects were found by placing a message on the usenet newsgroup sci.med.informatics asking for participants. This resulted in a wide range of participants, from the United States, Canada, and even a physician from France. There were no criteria for participation in the trial except that the individual had to be a medical doctor or in medical school. One respondent was a registered nurse. She was allowed to participate, but to help interpret results, all participants were grouped according to the amount of clinical training that they had received. Overall, over 80 packets, each containing 10 cases, were created and distributed. Of these, 22 packets were completed and returned. The large number of unreturned packets is due to several factors. One participant asked for 30 additional packets to be distributed to interns at the teaching hospital where he worked. Repeated queries were able to effect the return of the individual's packet, but the 30 additional ones were never returned. Similarly, a total of 14 packets were sent to a medical school where a colleague of Dr. Kohane was attending. Only 1 of those packets came back. Another participant distributed 10 packets to his colleagues, of which only 1 was returned. Of a total of 13 packets distributed to medical students, only 3 made it back.

After the packet distribution had begun, it was discovered that 3 of the 100 cases had typographical mistakes that were not caught earlier. They were removed from the trial, leaving 97 cases. Later, 2 more cases were removed from the trial because their medical records could not be located to obtain the medical record diagnoses. Thus, the final number of cases used in the trial was 95. Table 2: Medical Record Diagnoses of Trial Cases, lists the breakdown of the 95 cases.

Medical Record Diagnoses of Trial Cases

| Category Name | Description / Diagnosis | Number |
|---------------------------|---|--------|
| Normal - | Normal Growth, Early Puberty, Constitutional Delay of Puberty, Familial Short Stature | 50 |
| Normal - Other | Referred for non-growth problem | 18 |
| Precocious Puberty | Precocious Puberty | 6 |
| GH-Def and Hypothyroidism | Congenital Growth Hormone Deficient, Acquired Growth Hormone Deficient, Hypothyroidism | 11 |
| Complex Cases - | Multi-congenital abnormalities/Cancer | 8 |
| SB/ Turner's | Short Bone Syndrome, Turner's Syndrome | 2 |

Table 2: Medical Record Diagnoses of Trial Cases

3.1.5. The Task

Each of the participants was told that the cases which they were reviewing came from files from the endocrine clinic at the Boston Children's Hospital. They were also told that the data that they were presented with consisted of all the height, weight, and bone-age data available to the physician at the time that the child was referred. In addition, they were reminded that not all of the patients had growth disorders. They were asked not to discuss the case with others or "study" in preparation for participation, and they were told to spend the same amount of time that they would normally spend if asked in a clinical setting to give an opinion.

Each growth chart presented the data graphically and in tabular form (see Appendix B - Packet Directions / Samples to see a sample patient chart). At the bottom of each chart is a response area in which the subject was asked to do three things (See Figure 6). First, they had to decide whether to refer the child to the growth clinic. Then they were asked to give a preliminary diagnosis. Finally, if they felt that a referral was warranted, they were asked to choose a time point at which it would have been appropriate to refer the child, only having seen the patient's data up to that point. For example, if the subject felt that the data suggested that the child had a Short Bone Syndrome, the subject should then choose to refer the child to the clinic and place a check next to the Short Bone Syndrome / Turner's Syndrome / Hypochondroplasia diagnosis. Then, if the subject felt that the child's clinical measurements clearly showed a growth abnormality that should have been noted by the data point at age 6, then he/she would circle the 6 in the tabular listing of the clinical measurements adjacent to the graphical picture of the patient's growth chart.

Sample Human Subject Response

1. Based on the data presented, would you recommend:
 Approve referral to endocrine clinic Deny referral to endocrine clinic
2. Please place a checkmark next to exactly one congenital condition and any number of acquired conditions that you feel best describe the patient.

| Congenital Conditions | | Acquired Conditions |
|---|---|---|
| <input type="checkbox"/> Normal Growth | <input checked="" type="checkbox"/> Precocious Puberty | <input type="checkbox"/> Acquired Growth Hormone Deficiency |
| <input type="checkbox"/> Early Puberty | <input checked="" type="checkbox"/> Short Bone Syndrome / Turner's Syndrome / Hypochondroplasia | <input type="checkbox"/> Hypothyroidism |
| <input type="checkbox"/> Constitutional Delay | <input type="checkbox"/> Not Enough Information | <input type="checkbox"/> Obesity |
| <input type="checkbox"/> Congenital Growth Hormone Deficiency | | |

3. Please circle the age at which you feel the patient should be referred

Figure 6: Sample Human Subject Response

3.2. Gold-Standard - Medical Record and Experts

There were two gold-standards in this evaluation. The first gold-standard was the diagnosis written in the medical record of the patient. A second gold-standard was the evaluation of the patient by a pediatric endocrinologist.

The medical record diagnoses for the cases were first obtained from the on-line problem list of the patient. In approximately 75% of the cases, the problem list was empty so the most recent referral letter was scanned and any diagnoses made by the endocrinologist who saw the patient were accepted. A referral letter is the letter sent back to the pediatrician who referred the patient to the growth clinic. It contains the patient's history, findings, diagnoses, and other clinical information. Because the on-line problem list appeared so incomplete, the most recent referral letter was consulted for each and every patient, even if the problem list was not empty. The union of the diagnoses from the problem list and the referral letter was accepted as the correct diagnoses. There were no cases where the two sources were incompatible. Some of the patients had no data in the on-line medical record. In those cases, the physical medical record was used. In the end, 2 of the cases had incomplete medical records and were removed from the trial.

To obtain the answers for the second gold-standard, a pediatric endocrinologist was given the same sheet that was given to the human subjects. Four endocrinologists each saw one quarter of the approximately 100 cases. The endocrinologist that helped develop Trendx was **not** one of the four endocrinologists who provided the gold-standard. In a

similar fashion to the human subjects, the experts were asked to either recommend or deny a referral to the growth clinic and to circle the appropriate time of referral. However, instead of choosing one diagnosis, the experts were asked to rank up to three acceptable preliminary diagnoses. Figure 7 shows a sample gold-standard response for a patient.

Sample Gold-Standard Response

1. Based on the data presented, would you recommend:

Approve referral to endocrine clinic Deny referral to endocrine clinic

2. Please place a checkmark next to exactly one congenital condition and any number of acquired conditions that you feel best describe the patient.

| Congenital Conditions | | Acquired Conditions |
|--|---|---|
| <input type="checkbox"/> Normal Growth | <input checked="" type="checkbox"/> Precocious Puberty | <input type="checkbox"/> Acquired Growth Hormone Deficiency |
| <input type="checkbox"/> Early Puberty | <input checked="" type="checkbox"/> Short Bone Syndrome / Turner's Syndrome / Hypochondroplasia | <input checked="" type="checkbox"/> Hypothyroidism |
| <input checked="" type="checkbox"/> Constitutional Delay | <input type="checkbox"/> Not Enough Information | |
| <input checked="" type="checkbox"/> Congenital Growth Hormone Deficiency | | |

3. Please circle the age at which you feel the patient should be referred

Figure 7: Sample Gold-Standard Response

There are several reasons for using an expert opinion as a gold-standard in addition to the medical record gold-standard. As discussed in Section 3.4, expert opinion is an accepted gold-standard for trials of medical expert-systems because true gold-standards often do not exist.. This applied to our trial as well. Since some of the patients in our test set were referred for problems that were not related to growth, it would not be possible to arrive at the medical record diagnosis from only the height, weight, and bone-age data. Instead, we consider these cases “normal,” even though the medical record does not explicitly state that the child is normal. Moreover, the doctor in the growth clinic who saw the patient had much more information available to him or her. This included a complete physical exam as well as the ability to take more measurements and labs. In fact, the medical record diagnosis might not have been made until several visits after the first referral visit. This was too high a standard for any individual to be held against, especially considering the limited amount of data available. A final reason to use an expert opinion was the poor quality of the information available in the medical record. As noted, the problem lists were incomplete and the referral letters often mentioned the *possible* presence of other disorders that were never confirmed or denied.

3.3. Performance Measures

The first question which must be answered is, “Is it possible to make a decision about whether to refer a child based on only height, weight, and bone-age data?” To answer that question, we compare the recommendation of the pediatric endocrinologist to the diagnosis written in the patient’s chart.

To evaluate the performance of the test subjects, their decisions are compared to those of both gold-standards for all cases, and then for the set of cases in which there was consensus between the gold-standards. Finally, we use a scoring mechanism to rate the performance of each subject

The following sections describe the comparisons in more detail. Note that a child with a clinical diagnosis of normal growth, early puberty, constitutional delay of puberty, or familial short stature was considered normal and not in need of referral. In addition, any patient that was diagnosed with a disorder that would not affect the child’s growth, as determined by the expert who helped develop TrenDx, was also categorized as “normal.”

3.3.1. Comparison To Medical Record Diagnosis

First, the expert opinions, the diagnosis of TrenDx, and the diagnosis by the human physicians were all compared to the medical record diagnosis. In the most basic analysis, the decision of the subject about whether to refer the child was compared to the clinical outcome of the child (normal vs. abnormal). Then, for all of the abnormal patients, the preliminary diagnosis made by TrenDx and the human subjects was compared to the medical record diagnosis. All of these comparisons were performed for the human subjects as a group, and for the sub-populations formed by separating the physicians by the amount of training that they had received.

3.3.2. Comparison to Expert.

The decisions of the subjects were also compared to the recommendations of the pediatric endocrinologists. We performed the same comparisons that were used with the Medical Record Diagnosis. In addition, a scoring mechanism was devised that would allow us to simplify the comparison. Numeric scores allow the performance to be quickly summarized as well as grouped over many patients. However, any scoring mechanism has biases and weaknesses. (Hayes-Roth, Waterman and Lenat) note that:

Principle 1. Complex objects or processes cannot be evaluated by a single criterion or number.

Principle 2. The larger the number of distinct criteria evaluated or measurements taken, the more information will be available on which to base an overall evaluation.

3.3.3. Scoring Mechanism

Again, the decisions that the test subjects needed to make were:

- 1 - Refer or not to Refer
- 2 - Make a Preliminary Diagnosis
- 3 - Choose a Time to Refer.

Clearly, the decision to refer or not to refer was the most important. This is because if the patient did have a growth problem and was referred to a specialist, the specialist should be able to make the correct diagnosis. Giving the decision to refer a value of 5, the preliminary diagnosis a possible value of 3, and the timing of the referral a possible value of 2 gave each case a total possible score of 10.

The possible number of points for the preliminary diagnosis score was 3. Recall that the gold-standard expert ranked up to three preliminary congenital diagnoses, as well as checking off any number of the acquired conditions (See Figure 7). Since the number of acquired conditions that were chosen by the expert varied, the 3 points were divided as follows, depending on whether the expert felt that any acquired conditions were appropriate to note.

- **No Acquired Conditions Appropriate:** 3 points for choosing the top-ranked congenital diagnosis as the diagnosis, 2 points for choosing the second-ranked diagnosis, and 1 point for the third.

- **One or More Acquired Conditions Appropriate:** 1 point for choosing any of the acquired conditions. 2 points remaining for the congenital conditions: 2 points for the top-ranked congenital diagnosis, 1 point for the second, 0 points for the third.

To make the scoring system more concrete, let's match the sample human subject response in Figure 6 to the gold-standard response in Figure 7. First, note that both the human subject and the gold-standard felt that the patient should be referred. That gives the human subject 5 points out of 5 possible points. For the preliminary diagnosis, the gold-standard ranked 2 congenital disorders, Congenital Growth Hormone Deficiency and the Short Bone Syndrome / Turner's Syndrome / Hypochondroplasia combination. The gold-standard expert also noted that the patient was obese. This causes the scoring possibility to fall into the second of the two scoring categories listed above - "One or More Acquired

Conditions Appropriate.” The human subject did not note the acquired condition of obesity, and his/her choice of congenital conditions matches the 2nd choice of the gold-standard expert. According to the scoring scheme this scores 1 point out of 3 possible points.

The 2 remaining points, based on the timing of the referral, were also split - 2 points for choosing the referral at the same time as the gold-standard, and 1 point for coming within 1 data point, if it was within 2 years of the correct time. This scoring mechanism allowed subjects to score if the time at which they felt that the referral should have been made was “close enough” to that of the expert. To continue our example, assume that both subjects felt that the referral should have been made at the time that the patient was 6 years old. Thus, the subject scores 2 out of 2 possible points, for a total score of 8 out of 10 for this case. Note that the timing of the referral is not shown on the small portion of the response sheet that we have presented in Figure 6 or Figure 7. It can be seen in on a complete response sheet as shown in **Appendix B - Packet Directions / Samples** .

As a side-effect of the scoring scheme, the number of possible points became fewer than 10 if the Gold-Standard expert decided that the patient should not be referred, because the 2 points for the timing of the referral could not be scored. In those cases, the referral decision became worth 7 points to keep the total possible points the same for all cases.

Table of Results for a Sample Patient

| Sub: | Refer (y/n) | Ref Age | Early Pub | Norm | Cons Del. | Cong GH | Prec Pub | Short Bone | No. Info | Acq. GH | Hyp Thyr | Obes | Score |
|-------|----------------|------------|--------------|------|--------------|------------|-------------|---------------|-------------|------------|-------------|------|-------|
| MR | N | | | X | | | | | | | | X | 1 |
| Exp | Y | 4.25 | 2 | | | | 1 | | | | | X | N/A |
| TrDx | N | | X | | | | | | | | | X | 2 |
| Sub1 | Y | 4.25 | | X | | | | | | | | X | 9 |
| Sub21 | N | | | X | | | | | | | | X | 1 |
| Sub60 | Y | 5.5 | | | | | X | | | | | X | 9 |

Table 3: Table of Results for a Sample Patient

Table 3 is an example of a table of results for a sample patient. The table lists the complete set of answers for a patient, including the medical record diagnosis, the opinions of the gold-standard expert, the decisions of TrenDx, and all the responses of the human

subjects. The complete set of results is listed in **Appendix A - Patient and Subject Result Tables**. The columns are:

- Sub: The subject giving the answers.
- Refer (y/n): Whether the subject decides to refer the patient
- Ref Age: If referral was recommended, the age of the child that the referral should have been made
- Early Pub: Early Puberty
- Normal: Normal Growth
- Cons Del.: Constitutional Delay of Growth
- Cong GH: Congenital Growth Hormone Deficiency
- Prec Pub: Precocious Puberty
- Short Bone: Short Bone Syndrome/Turner's Syndrome
- No. Info: Not enough information
- Acq. GH: Acquired Growth Hormone Deficiency
- Hyp Thy: Hypothyroidism
- Obes: Obesity
- Score 1: Score relative to answers provided by Gold-Standard Expert.

Within the Sub column, the rows are:

- MR: Medical Record Gold-Standard
- Exp: Gold-Standard Expert
- TrDx: TrenDx answers
- Sub##: Human Subject number ## answers

To illustrate the scoring mechanism again, the score for TrenDx is calculated as follows. The decision to refer is Exp- Y, TrDx- N, so TrenDx scores 0 points for the decision not to refer the child. That automatically prevents TrenDx from scoring in regards to the timing of the referral, since TrenDx did not refer the child. In the area of the preliminary diagnosis, TrenDx gets 1 point for choosing the acquired condition of Obesity and 1 point for choosing the second-ranked condition of Early Puberty. Thus TrenDx scores 2 points in comparison to the Gold-Standard Expert.

3.3.4. Other Comparisons

In the evaluation of four diagnostic decision-support systems, Berner uses several measures of performance based on a consensus of the programs being tested (Berner et al.). We made similar measurements by noting the number of cases in which more than one human subject reviewed the case and in which the referral decisions were in agreement. Out of these cases, we looked at several statistics such as the number of cases in which both gold-standards agreed and the number of cases in which any group made a singular decision. A singular decision is one in which one group makes a decision and every other group makes the opposite decision (eg. chose to refer while all other groups chose not to).

3.4. Comparison to Other Evaluations

Since evaluations are an integral part of the development of decision-support systems, or expert-systems, many different types of evaluations have been performed (Berner et al.; Feldman and Barnett; Heckerman and Nathwani; Miller, Pople and Myers). Most of these evaluations attempt to measure the performance of the system on some number of cases. Forsythe argues that performance should not be the only aspect of a system that is measured (Forsythe and Buchanan). This is especially true of field trials of systems that are very advanced along the development cycle.

In evaluations of expert systems, several ways of obtaining a gold-standard or evaluating the answers produced by the system have been devised. The gold-standard is usually either the “real” answer, such as the correct diagnosis as confirmed by laboratory studies, or the opinion of one or more experts in the domain. In some cases, the expert that helped develop the system is also involved in the evaluation (Heckerman and Nathwani) introducing bias into the evaluation. Often, the system is the only thing being tested and its answers are evaluated by an expert and given either a subjective rating or a quantitative rating. Sometimes both the system and other physicians, who are not considered experts in the particular domain, are both evaluated and their performance is compared.

Several systems which try to cover a wide domain, such as the entire field of internal medicine, try to produce a ranked differential diagnosis list (Bankowitz, Lave and McNeil; Feldman and Barnett; Miller, Pople and Myers). In these cases, evaluation is more complicated because the goal of the program may not be to just suggest the most likely answer, but to also stimulate the user by suggesting rare conditions. In these cases, more complex measures are devised to represent favorable and unfavorable traits. These generally involve counting disagreements and agreements between all of the participants. Specifically, not suggesting a diagnosis that every other participant suggested and being the only participant to suggest a particular diagnosis are two unfavorable characteristics.

In terms of the patient cases, our evaluation of TrenDx differs from many other evaluations because most of our patients are normal. Again, the cases were chosen serially with some screening criteria and then “cleaned” by collecting all the measurement data and by removing patients whose medical record contained unreadable measurements or had similar problems. While this introduces bias into the evaluation by limiting the scope of the trial, it also simplifies the evaluation by making it easy to categorize patients and avoid problems with missing data or cases in which the child somehow loses 10 centimeters in height between one visit and the next. As noted before, these cases were simply removed

from the trial. In a field trial of the program, TrenDx would have to be programmed to deal with poor data.

4. TrendX Development

The development of TrendX in preparation for the trial can be broken down into three areas: programming new features and fixing bugs in the inference engine, modeling the processes that we wanted to monitor in order to get trend-templates, and knowledge-engineering the trend-templates to achieve improved performance.

4.1. Programming

The state of TrendX at the conclusion of Haimowitz's thesis was poor. In fact, in running the trial discussed in his thesis (Section 5.1 Haimowitz), he was hampered by the fact that it took several hours to run a single patient. This was a direct result of the speed of the TUP functions and the large number of unconstrained intervals in his trend-templates. There were also several bugs in the inference engine that caused Now-Based trend-templates to crash.

Together, Haimowitz and I were able to fix the problems with Now-Based trend-templates. This improvement was essential in order to use several of the new trend-templates such as the trend-template for Obesity and Acquired Hypothyroidism. I also worked with Kohane to improve the speed of TUP calculations, since they constituted the majority of the time it took to process a patient.

The modeling language was improved by adding the ability to specify ranges on constant and linear constraints. For example, a user could specify that a parameter should be constant and that the parameter's value should be within the interval -5 to 5.

4.2. Modeling Growth

4.2.1. Standards

Several standards are routinely used to measure the growth of a child. The National Center for Health Statistics, or NCHS, has produced cross-sectional population curves for growth of both boys and girls up to age 18. Tanner and Davies produced longitudinal standards for height and height velocities that are more appropriate to use when following the growth of a child once puberty has begun (Tanner and Davies). These standards include curves for children in whom puberty occurs at the average time, as well as 2 standard deviations early and late. To compare the patient to a standard, we calculate the Z-

score for that measurement. Again, the Z-score represents the number of standard deviations from the mean.

During the previous development of thesis-TrenDx, it was noted that a clinically relevant piece of information was a measure of the weight of the child relative to his or her height. Using the weight and height curves developed by the NCHS, we created a measurement known as Build to represent the stockiness of the child. This Build measurement was chosen over another measurement known as the Body-Mass Index, or BMI, because the BMI was developed as a measurement of adult obesity. Several studies have shown that BMI is not a good indicator of obesity, or body build, in children (Roche et al.; Roland-Cachera et al.). During the current trial of TrenDx, the use of the Build measurement over the BMI was reconfirmed by showing that the BMI for fictitious patients who follow different percentile curves is not consistent over the longitudinal growth of a child. By definition, the Build measurement is consistent over the longitudinal growth of a child who follows percentile curves.

The pubertal development of children is measured on what is known as a Tanner Scale(Tanner). The scale ranges between 1 and 5, where 1 signifies no pubertal development and 5 signifies full pubertal maturity. A score of 2 signifies the beginning of puberty, and progression to each successive stage generally occurs within 6 months to 1.5 years. There are three Tanner scores - for pubic hair, breast, and penis development. Generally, only pubic hair and penis development scores are appropriate for boys, while pubic hair and breast development scores are appropriate for girls. In this trial, since the pubertal development information was not available, it was not used in the trend-templates actually used in the trial. However, pubertal development constraints should certainly be included in any model of pediatric growth.

The skeletal development of a child is measured by taking an X-ray of the left hand and wrist and comparing to published standards. The skeletal age, which we term bone-age, is very valuable in determining the growth state of a child. In fact, Kaplan uses bone age to differentiate between primary and secondary disturbances of growth(Kaplan). In essence, a short child who has a bone-age that is younger than his or her chronological age has the potential to continue growing when other children have stopped. This means that the child may “catch up” to his or her peers. The disorders which result in this type of growth delay are termed secondary disturbances of growth. Constitutional Delay of Puberty and Congenital Growth Hormone Deficiency are both secondary disturbances of growth. In comparison, a short child that does not have a delay in his or her skeletal growth does not have this same potential to “catch up.” This characterizes a primary

disturbance of growth, such as a skeletal dysplasia - which we group under short-bone syndrome in our set of trend-templates.

4.2.2. Description of Growth States

The following is a characterization of all the process states, or growth conditions, that we chose to model. Collectively they are referred to as the set of disorders, even though they include normal growth. These characterizations were taken from consulting the expert as well as from descriptions listed in medical texts (Becker; Kaplan; Roche et al.; Roland-Cachera et al.; Tanner; Tanner and Davies). They are presented to aid the reader in understanding the trial and are only grossly correct from a medical standpoint. All of the trend-templates that we developed are listed in Appendix C - Trend Template LISP Code.

NORMAL GROWTH

Normal infants generally develop with consistent height and weight Z-scores up until sometime between ages 2 and 3. Sometime between 2 and 3, they ‘establish centiles,’ meaning that they naturally fall into a certain percentile channel that they stay within until they stop growing. The exact time of the onset of puberty is generally accepted to be around the age of 9 to 12 years in females and 10 to 13 years of age in males. Females continue growing slowly almost to the age of 17 whereas males often continue growing beyond age 18.

CONSTITUTIONAL DELAY OF PUBERTY

In children with Constitutional Delay, the skeletal development of the child is generally delayed in proportion to the delay in development of their height. This delay can be up to two years or more in extreme cases. Consequently, puberty occurs somewhat later, and they continue to grow for a longer period of time. Final adult height is equivalent to children with Normal Growth.

EARLY PUBERTY

Children with Early Puberty are characterized as having an advanced skeletal age. Puberty occurs earlier, as does cessation of growth. Final adult height is still equivalent to children with Normal Growth.

Being described as early or delayed signifies that the child is early in relation to the normal population. There is no agreed-upon definition of what is early and what is not. While both Constitutional Delay and Early Puberty are considered benign, the extent of the delay or acceleration may cause an expert to consider it differently.

CONGENITAL GROWTH HORMONE DEFICIENCY

Children suffering from Congenital GH Deficiency have several traits that we can model. They are generally significantly short, directly due to an inability to produce or respond to growth hormone. In addition, the extent of their skeletal and sexual delay is much greater than in Constitutional Delay of Puberty. Often these children have sexual infantilism.

SHORT BONE SYNDROME / TURNER'S SYNDROME

We use the category Short Bone Syndrome/Turner's Syndrome to describe a large class of disorders which we feel can not be differentiated by only height, weight, pubertal, and bone-age data. We characterize the group as being even shorter than Congenital GH Deficient children, with very little delay in skeletal age.

PRECOCIOUS PUBERTY

True precocious puberty describes a condition in which children develop sexual characteristics at an extremely early age. As one would expect, skeletal age is more advanced than chronological age, and the children are often taller than their peers because of their advanced development.

ACQUIRED GROWTH HORMONE DEFICIENCY

Acquired Growth Hormone Deficiency is a condition that is acquired at some time during the patient's life. Prior to this time, he or she may be perfectly normal. This condition is often secondary to some other problem, such as a craniopharyngioma. The effects of this disease are a marked deceleration of growth at the onset of the deficiency. Experts often talk about a child "falling off" the growth curve. However, the rate of deceleration is often variable since it is affected by the source and extent of the problem. This makes it difficult to model.

CONGENITAL AND ACQUIRED HYPOTHYROIDISM

There are many clinical manifestations of Acquired Hypothyroidism such as lethargy, decreased appetite, and other findings which are not available to us. However, Kaplan notes that "...the most important sign of acquired hypothyroidism in childhood is growth failure."(Kaplan) Height moves in a progressive downward deviation and weight tends to increase modestly. Skeletal development is delayed in proportion to height age. Kaplan also notes that these are characteristic of hypopituitarism (Growth Hormone

Deficiency), and consequently it is difficult to differentiate between the two from the data available to us.

OBESITY and MALNUTRITION

These conditions were included because the growth clinic is often sent referrals for children with extreme weight problems. As noted, weight, or more precisely build, is clinically significant in several of the disorders we are trying to diagnose. We cannot actually capture obesity information from the data that we have, since we cannot differentiate between a muscular individual and an obese individual from only height and weight data.

It is very difficult to differentiate between Growth-Hormone Deficiency and Hypothyroidism from the available data. Moreover, it is also difficult to differentiate between Growth-Hormone Deficiency and Hypothyroidism, and the Short Bone Syndrome except by knowing Bone Age and, to a lesser extent, Pubertal Stage data. However, both of these data types are not routinely collected and thus are not present in very many of the test cases. Consequently, Growth Hormone Deficiency and Hypothyroidism were considered the same diagnosis in both the creation of our trend-templates and in scoring Trendx and the human subjects.

4.2.3. Modeling

The trend-templates used to model each of the disorders presented in the previous section were designed based on consultations with the expert, reading medical texts, and adapting the trend-templates used in thesis-Trendx. The entire set of trend-templates is listed in Appendix C - Trend Template LISP Code.

The new (current) and old (thesis-TrenDx) trend-templates for normal growth in

Current Trend Template for Normal Male Growth

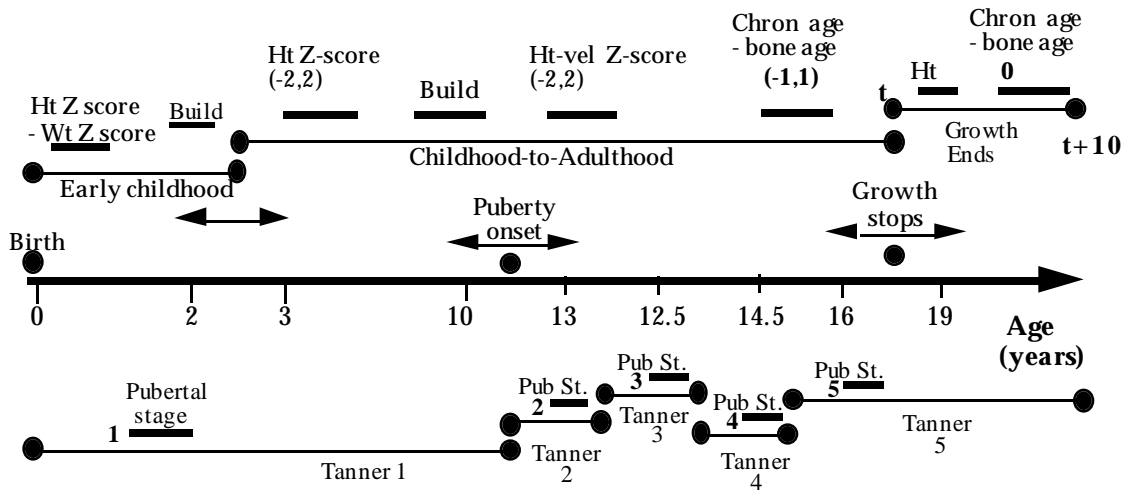


Figure 8: Current Trend-Template for Normal Male Growth

males are shown in Figure 8 and Figure 9 for easy comparison. The new trend-template differs from the previous trend-template in several ways. One of the most notable is the reduction of Intervals 2 through 4 in Figure 9 to a single interval in Figure 8. Another is the absence of the second order value constraints, and a third is the addition of numeric values as part of the constraints.

Haimowitz notes that:

First, in order to insure reliable value constraint matches, TrenDx should assign at least three or four data points per trend template interval. When modeling trends in sparse data sets, a knowledge engineer should minimize the number of disjoint intervals constraining a particular parameter. If trend templates contain too many such intervals, TrenDx may find a low-scoring or trivial match by assigning only one or two data points to each interval. (Section 5.1.5 Haimowitz)

This is especially true for the domain of pediatric growth monitoring, when measurements are not necessarily taken with any regularity. In fact, because data in growth monitoring is often sparse, each individual data point takes on much more significance. This is in contrast to some domains where data collection occurs at a high frequency, such as ICU data monitoring. In those domains, individual data points may have little significance compared to the overall trend in the data.

Previous Trend-Template for Normal Male Growth

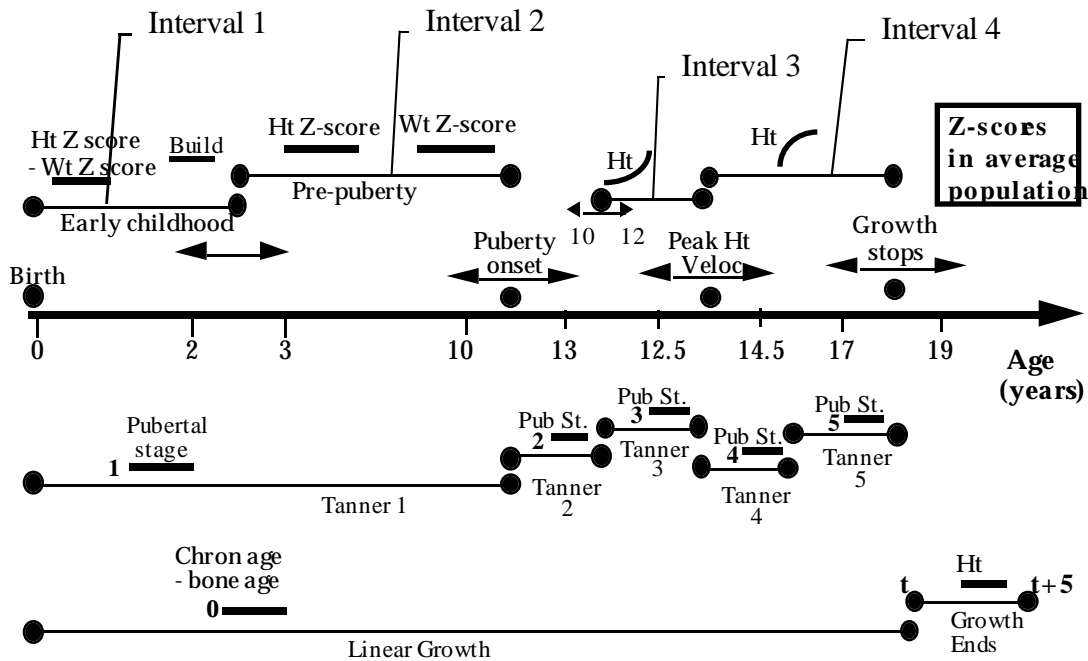


Figure 9: Previous Trend-Template for Normal Male Growth(Haimowitz)

Note that in the previous trend-template (Figure 9), there are four different intervals that constrain the patient's height or height Z-score (Intervals 1-4). Depending on the value constraint, each interval needed two or three data points before a non-trivial score could be assigned. This problem is compounded by the ability to specify variable value constraints. For example, a user could specify that a particular parameter should be constant at some unknown value. In that case, a single value in the interval would match trivially, since the value constraint would match a single value to a wildcard and get no error. Moreover, the presence of multiple intervals with uncertain endpoints increases the computational complexity of matching data to trend-templates. Each data point that falls into the uncertainty range of an interval endpoint causes *TrenDx* to branch and create two new contexts - one in which the data belongs to the interval and one in which it doesn't. For both of the new hypotheses, all the data must be matched to the trend-template and both of the hypotheses kept until one or both are pruned by the beam-search.

There is also a discontinuity between Interval 2 and Interval 3. Since the temporal constraint allowed the Begin point of Interval 3 to vary, *TrenDx* would branch when there was a data point between the ages of 10-12. One branch would include the data point in the interval and the other would not. The beam search would then choose the second

hypothesis over the first because in the second temporal context, the data point does not fall into any time interval and is therefore not matched against any constraints. Thus, it contributes no error-score to the score of the second hypothesis. In essence, TrendX was ‘throwing away’ data.

Using the lessons learned from these mistakes, new models of the growth states were created and trend-templates were written from these models. These trend-templates were further refined during iterative sessions of knowledge-engineering, discussed in Section 4.3.

Haimowitz suggests that “A trend template should only include information that will distinguish it from competing trend templates.”(Section 6.5.1 Haimowitz) This suggestion is based on the fact that having the same information in each trend-template will contribute the same amount of error to each trend-template. While this is true, it ignores the fact that knowing that a process matches to all the trend-templates poorly is important information in itself. In fact, in a domain where all of the possible states are not known (most real-world domains), using this information suggests that the process is in some unmodelled state, assuming that the models are correct. The principle is amended as follows:

A trend-template should include as much information as practical, even information that does not distinguish it from competing trend-templates.

It is important to examine exactly what knowledge is expressed while modeling with TrendX. Unlike many other expert-systems, such as INTERNIST-1 (Miller, Pople and Myers) or PATHFINDER (Heckerman and Nathwani), there is no implicit or explicit mention of the probability of the occurrence of each growth disorder or some type of evoking strength of some particular symptom. There are some implicit probabilities embedded in the performance of the program, as well as assumptions about the distributions in height and weight that are used to make Z-score calculations; however, there are no statements equivalent to “Short Bone Syndrome occurs in 1 out of every 35,000 individuals,” or “A Bone Age 1 year behind chronological age indicates the presence of Congenital Growth Hormone Deficiency with a probability of 0.35.” While it is true that knowledge of prior probabilities of diseases and findings is important and useful for diagnosis, the absence of these probabilities allows TrendX to model processes in which these probabilities are not known.

In modeling the Acquired Growth-Hormone Deficiency trend-template, another lesson was learned. The first attempt to model the disorder only included an interval that extended back 2 to 3 years from the most recent data point. TrendX created alternate

temporal worlds for each data point that might have been included in the interval. Since each hypothesis received an error-score and the hypotheses were pruned using a beam search, it was most likely that hypothesis in which the sole interval contained the fewest number of points would score the best. Again, TrenDx was “throwing away” data. Thus, the trend-template was extended by adding an interval that represents the time before the onset of the growth-hormone deficiency. By defining the intervals to be consecutive, a data point must fall into one of the intervals and no data is thrown away. The best hypothesis becomes the one with the best onset time instead of the one with the shortest time interval. This is the same problem that was encountered with the non-consecutive intervals in the previous trend-templates used in Haimowitz’s thesis trial.

Overall, it is difficult to characterize abnormal growth. Medicine does not understand every disorder that afflicts man. Even when the disorder is known and has been characterized, its effects on each individual can vary in presence or absence of symptoms and the severity of those symptoms. The possibility of multiple disorders being present and interacting further complicates matters. Because of all these factors, it is imperative to model the normal state as well as possible.

4.3. Trend Template Refinement

Once the broad outline of the trend-templates were generated for each of the disorders, they were refined to achieve the desired performance. Many of the age ranges from the previous trend-templates were used, and the improvements mentioned in Section 4.2.3, such as reducing the number of intervals, were made. In addition, the values for parameter error-function models and thresholds had to be established.

4.3.1. Establishing Performance Goals and Training Sets

The development and refinement of trend-templates is an iterative process. A user must define a trend-template and then process the data for a particular patient and see how the trend-template performs. If the performance does not match the desired performance, then the trend-template must be refined and the process must be repeated.

A set of performance goals were defined to give a concrete definition of when the trend-templates were acceptable. The performance goals consisted mainly of specifying that the trend-templates should score well or poorly on a training set of fake patients with stereotypical patterns of growth. The training set also consisted of some actual patients such as the cases used in earlier trials of TrenDx. The existence of the training set was

important to ensure that new additions to the knowledge base did not interact with the previous information in unexpected ways.

4.3.2. Error-Function Models

At the conclusion of Haimowitz's thesis, Trendx was somewhat limited in the expressive capabilities of the error-function of a value constraint. It allowed the user to specify an expected value of a parameter, such as (constant 5), but the ability to specify a normal range was absent. After this ability was incorporated, it was discovered that it still did not capture the knowledge that the expert wanted to express. The expert wanted to say, "the value should be in the range of A to B. If the value is somewhere near the middle, that's good. If the value is near the endpoints, then the value does not match the trend-template very well. If it is outside the range, the trend-template should score very badly." Constraint-Based Trendx, discussed in Section 2.3, was only able to distinguish between the first and last conditions while Regression-Based Trendx gave the user the ability to express the first and second conditions. Looking at the definition of the value constraint as a composition of functions, *the limitations of the second component can be overcome by making the first component more complex*. As a simple example, the performance of Constraint-Based Trendx could be imitated by the Regression-Based Trendx as shown in Figure 10.

Comparison of Simple and Compound Value Constraints to Constraint-Based TrendDx

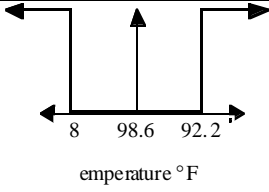
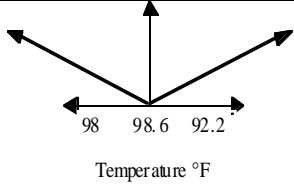
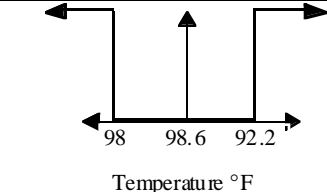
| | Constraint-Based TrendDx | Regression-Based TrendDx | |
|---------------------------------|--|--|---|
| | Normal Value Constraint | Simple Value Constraint | Compound Value Constraint |
| Parameter | Temperature | Temperature | Function (Temperature) if Temperature < 98.0°F return 1 if Temperature > 99.2°F return 1 else return 0 |
| Model | 98.0° to 99.2°F | (Constant 98.6°F) | (Constant 0) |
| Resulting Error- Function |  |  |  |

Figure 10: Comparison of Simple and Compound Value Constraints to Constraint-Based TrendDx

In fact, almost any error-function can be created with the use of various parameter functions. A few possible ones are shown in Figure 11.

By using a function in the parameter component of the value constraint, many different composite error-functions could be created. For many of the value constraints used in the final trend-templates, the composite error-function IV from Figure 11 was used.

It is arguable that these permutations on the first component of the value constraint are unnecessary and that instead, programming changes should be made within TrendDx to allow the expression of the more commonly used composite error-functions. The same results could have been achieved by programming new functions into the second component of the value constraint. However, time did not permit this and the original

programmer of Trendx was unavailable to help complete the task. Furthermore, each user of Trendx should not be expected to go into the internal code of Trendx to program new error-functions.

Possible Composite Error-Functions

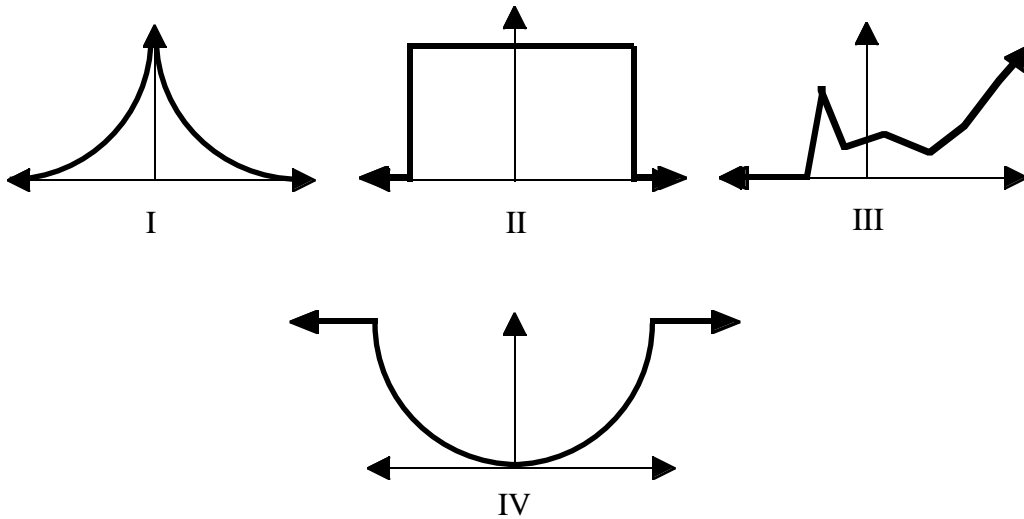


Figure 11: Possible Composite Error-Functions

We chose to use an error-function that looks like Error Function IV in Figure 11. This error-function represents the way the expert compares a value to its expected range. If the value is somewhere near the middle, the error is small. However, as the value moves away from the middle, the error increases gradually. Near the boundaries, the error is large. Since it is continuous and fairly smooth, this error-function avoids the brittle behavior that the Constraint-Based Trendx suffered from. A value just below the boundary scores almost as poorly as a value beyond the boundary.

4.3.3. Residual Mean Square Error vs. MAPE

Trendx allows the option of using either the Residual Mean Square Error or the Mean Absolute Percent Error when matching data to value constraints. The Residual Mean Square Error is defined in Equation 1.

$$\frac{\sum_t (\text{Expected}_t - \text{Actual}_t)^2}{\text{Degrees of Freedom}}$$

Equation 1: Residual Mean Square Error

Haimowitz used the Mean Absolute Percent Error, or MAPE, to calculate the match of data to a value constraint. MAPE is defined in Equation 2.

$$\frac{\sum_t \left| \frac{\text{Expected}_t - \text{Actual}_t}{\text{Expected}_t} \right|}{\text{Degrees of Freedom}}$$

Equation 2: Mean Absolute Percent Error

The ability to use the MAPE was added because it is useful for comparing the goodness of fit between models of variables with different scales. For example, a constraint with an expected value of 100 and an actual value of 110 would be off by 10, but really only deviates from the expected value by 10%. In comparison, a constraint with an expected value of 10 and an actual value of 20 is still off by 10, but deviates from the expected value by 100%.

In using the MAPE, an unexpected problem arose because the expected value of many of the constraints is 0, and division by 0 is undefined. There were two possible solutions to the problem. One solution would be to add some offset to both the expected and actual values and then perform the calculation. However, that solution causes the error to be affected by the size of the offset. The second solution was to use the Residual Mean Square Error. The weakness of the residual error, matching to variables of different scales, was minimized by keeping the range of values of all variables approximately equal and by using the error-function IV shown in Figure 11.

4.3.4. Thresholds for Triggering

Recall that Trendx calculates an error score for each hypothesis of each trend-template. Thus, the hypothesis that matches the data the best is the one with the lowest error score. The decision to refer is based on the assumption that the trend-templates for Normal Growth, Constitutional Delay of Puberty, and Early Puberty are considered “normal” and do not require a referral to the growth clinic. Often, all error-scores are trivially small because only a few data points have been processed or because data points match trivially to certain value constraints. To avoid triggering a referral at these times, a threshold score was used. It was similar to the one described in the previous trial of

TrenDx. When the error-scores for all the “normal” trend-template hypotheses scored extremely high for a single data point, or somewhat high for two consecutive points, the patient was considered worthy of referral. Then, the best-scoring “abnormal” trend-template was taken as the diagnosis.

The triggering mechanism used to evaluate thesis-TrenDx (Haimowitz) was revised for this trial. Instead of triggering when the abnormal hypotheses scored a certain amount better than the normal hypotheses, we chose to trigger whenever all the normal hypotheses score worse than a certain threshold. We kept the use of the union of the two triggering styles: high single-point thresholds and lower consecutive-point thresholds (Section 2.6).

The threshold values were obtained from an analysis of the error-scores produced by processing 20 normal patient cases. The cases consisted of 10 male and 10 female patients, each with at least three data points. Of these cases, at least 3 of the patients of each gender had data points from infancy. The lowest error-score of the three “normal” trend-templates at each data point was analyzed. The mean and standard deviation were 0.19 and 0.08, respectively. We chose to set the threshold for the single-point triggering at 2 standard deviations from the mean, and the threshold for the consecutive-point triggering at 1.5 standard deviations from the mean. Thus, if a patient’s lowest error-score for all of the three “normal” trend-template hypotheses was above 0.35 at any time point, or above 0.31 for two consecutive time points, the patient was considered abnormal and a referral was triggered. The lowest non-trivial trend-template was then taken as the preliminary diagnosis. In some cases, most notably in infants, none of the trend-templates scored better than any of the others. In that case, we considered TrenDx unable to make a diagnosis, representing the “Not enough information” diagnosis. In addition, the Obesity acquired condition was considered “checked off” if the trend-template for Obesity scored better than the trend-template for normal-build at any time.

The results obtained by changing the triggering thresholds were also calculated to help justify the triggering mechanism. Recall that one of the reasons for the use of the union of the two triggering mechanisms was to reduce the brittle nature of thresholds, where a value just below the threshold would not trigger, while a value a little bit higher would trigger. To help understand whether this goal was achieved and whether the triggering threshold values that were chosen were appropriate, the effects of raising and lowering the threshold values were examined.

During the processing of the patients by Trendx, it was noticed that all hypotheses frequently scored very poorly on the first non-trivial ($\neq 0$) error-score. Specifically, in many cases the error-scores for a patient's first data point would be very high, then drop to below the triggering level. The reason for this was because some constraints require a certain number of data points before a meaningful match can be made. For example, if an interval had 2 constraints, one constraining a parameter to be equal to 0 and a second which constrained the same parameter to be linear with a slope of 0, then any single data point would match to the first constraint, but would match trivially to the second. Then, when the next data point was examined, Trendx would be able to generate an error-score for the second constraint. Since the overall error-score for the hypothesis is a weighted average of all the error-scores, the overall error-score could change dramatically between the first few data points. Therefore, we also looked at the effects of ignoring the first non-trivial error-score.

5. Results

Section 5.1 presents the results of comparing the decisions of the pediatric endocrinologists to the medical record diagnoses. First, the expert's referral decision is compared to the clinical outcome of the patient. Then, we categorized all abnormal patients by their disorder and we looked at both the percentage of patients that were referred, and the percentage of cases in which the preliminary diagnosis was correct.

Section 5.2 presents the results of comparing the decisions of the test subjects to the medical record diagnoses. We perform the same comparisons listed above, and additionally separate the human subjects by the amount of training that they have received.

In section 5.3, we compare the test subjects to the expert pediatric endocrinologists. We also look at the referral decisions for the set of cases in which the expert and the medical record both indicate that a referral is necessary.

Section 5.4 lists the results of applying our scoring mechanism to all test subjects, and Section 5.5 shows the results obtained by looking at "singular" decisions. Singular decisions are cases in which the recommendation of one of the four decision-making groups, the medical record, the human experts, the human subjects, and TrenDx, differed from the other three.

The performance obtained by using different threshold triggering mechanisms is explored in section 5.6. Finally, the timing of the decisions made by the test subjects is presented in section 5.7.

Of the 95 cases used in the trial, there were 68 normal patients and 27 abnormal patients. Table 4: Medical Record Diagnoses of Trial Cases, shows the number of cases in each population.

Medical Record Diagnoses of Trial Cases

| Category Name | Description / Diagnosis | Number |
|---------------------------|---|--------|
| Normal - | Normal Growth, Early Puberty, Constitutional Delay of Puberty, Familial Short Stature | 50 |
| Normal - Other | Referred for non-growth problem | 18 |
| Precocious Puberty | Precocious Puberty | 6 |
| GH-Def and Hypothyroidism | Congenital Growth Hormone Deficient, Acquired Growth Hormone Deficient, Hypothyroidism | 11 |
| Complex Cases - | Multi-congenital abnormalities/Cancer | 8 |
| SB/ Turner's | Short Bone Syndrome, Turner's Syndrome | 2 |

Table 4: Medical Record Diagnoses of Trial Cases

We used the following measurements in our comparisons to the Medical Record gold-standard:

- Sensitivity: (Patient is abnormal and referral approved) /
(Total # of abnormal patients)
- Specificity: (Patient is normal and referral denied) /
(Total # of normal patients)

Fisher's Exact Test was performed to obtain exact Chi-Square Test values. This was used to test for statistically significant correlations between the referral decisions of a test subject (refer / no referral) and the clinical outcome of the patient (normal/abnormal).

Note that sensitivity and specificity do not have their usual connotations in this trial because the source of the test cases is a population with a large percentage of abnormal children.

5.1. Expert vs. Medical Record

5.1.1. Referral Decision

Table 5 shows the results of the comparison of referral decisions between the expert pediatric endocrinologists and the medical record diagnosis. The expert referral decisions have a high sensitivity and low specificity. The P-value was not statistically significant at $\alpha=0.05$.

Expert Decision vs. Medical Record Diagnosis

| Decision | Abnormal | Normal | Total |
|-------------|----------|--------|-------|
| Refer | 19 | 40 | 59 |
| No Referral | 8 | 28 | 36 |
| Total | 27 | 68 | 95 |

Table 5: Expert Decision to Refer vs. Medical Record Diagnosis

Sensitivity: $19/27 = 0.70$

Specificity: $28/68 = 0.41$

Fisher's Exact Test - $P=0.3534$

5.1.2. Preliminary Diagnosis of Disorder Populations

Table 6 lists the referral and preliminary diagnosis performance of the experts, using the medical record as the gold-standard. This measurement compares the preliminary diagnosis given by the expert to the diagnosis written in the patient chart, over all the abnormal patients. Most of the columns are self-explanatory. There are two exceptions. The column labeled '#' represents the total number of patients with that disorder for whom a decision was made. This column total is not consistent between the tables for the experts, human subjects, and TrenDx. This is because the endocrinologists and TrenDx saw each patient exactly once, but both the Pre-Residency subjects and the Post-Residency subjects saw different patients, causing some patients to be evaluated more than others. The column labeled 'Correct Dx% (of Refs)' is the percentage of cases that had the correct diagnosis, out of the population of patients that were referred. It gives an indication of how often the diagnosis was correct, given that the subject felt that the patient had some kind of abnormality. For the 'Correct Dx%(of Refs)' Total calculation, the Cancer/Complex cases are not included.

Disorder Population Referral and Diagnosis Results Expert vs. Medical Record

| Disorder Sub Population | # | No Refer | Refer | Correct Ref % | Correct Dx | Correct Dx % | Correct Dx % (of Refs) ¹ |
|-------------------------|----|----------|-------|---------------|------------|--------------|-------------------------------------|
| Precocious Pub | 6 | 1 | 5 | 83.3% | 3 | 50.0% | 60.0% |
| GH-Def / Hypothyroid | 11 | 3 | 8 | 72.7% | 7 | 63.6% | 87.5% |
| Cancer / Complex | 8 | 2 | 6 | 75.0% | N/A | N/A | N/A |
| Short Bone / Turner's | 2 | 2 | 0 | 0% | 0 | 0% | 0% |
| Total | 27 | 8 | 19 | 70.4% | 10/19 | 52.6% | 76.9% |

Table 6: Disorder Population Referral and Diagnosis, Expert vs. Medical Record

The experts performed very well, having both a high Correct Referral % and the highest Correct Dx percentages (as compared to the test subjects). However, they completely missed both Short Bone/Turner's cases.

5.1.3. Summary

In summary, the experts had a high sensitivity (0.70), a low specificity (0.41), and a very high correct preliminary diagnosis percentage (76.9%). Their referral decisions did not correlate with the clinical outcome of the patient at a statistically significant level.

5.2. Test Subjects vs. Medical Record

5.2.1. Referral Decision

Table 7 through Table 10 list the results of the comparing the referral decisions made by TrenDx and the human subjects to the medical record diagnosis. The human subjects are divided into two groups - those who have completed a medical residency and those who have not.

Comparing the performance of TrenDx to the set of all the physicians, the sensitivities were the same (0.59), but the human subjects had a higher specificity value (0.53 vs. 0.47). When the participants were separated by training group, the Pre-

¹ Correct Dx% of refs does not include Cancer/Complex cases. Neither does Total.

Residency group had the highest sensitivity value (0.765) and a low specificity (0.47). The Post-Residency group had the lowest sensitivity value (0.52), and the highest specificity (0.54). None of the results were statistically significant at $\alpha=0.05$.

TrenDx Decision vs. Medical Record Diagnosis

| Decision | Abnormal | Normal | Total |
|-------------|----------|--------|-----------------|
| Refer | 16 | 36 | 52 |
| No Referral | 11 | 32 | 43 |
| Total | 27 | 68 | 95 ² |

Table 7: TrenDx Decision vs. Medical Record Diagnosis

Sensitivity: $16/27 = 0.59$

Specificity: $32/68 = 0.47$

Fisher's Exact Test - $P=0.6512$

All Physicians vs. Medical Record Diagnosis

| Decision | Abnormal | Normal | Total |
|-------------|----------|--------|-------|
| Refer | 35 | 75 | 110 |
| No Referral | 24 | 83 | 107 |
| Total | 59 | 158 | 217 |

Table 8: All Physicians vs. Medical Record Diagnosis

Sensitivity: $35/59 = 0.59$

Specificity: $83/158 = 0.53$

Fisher's Exact Test - $P=0.1296$

² TrenDx crashed while processing 3 of the patients. 1 was abnormal and 2 were normal. We assume that it got the wrong answer (referred normal patients and did not refer abnormal patients).

Pre-Residency Subjects vs. Medical Record Diagnosis

| Decision | Abnormal | Normal | Total |
|-------------|----------|--------|-------|
| Refer | 13 | 17 | 30 |
| No Referral | 4 | 15 | 19 |
| Total | 17 | 32 | 49 |

Table 9: Pre-Residency Subjects vs. Medical Record Diagnosis

Sensitivity: $13/17 = 0.76$

Specificity: $15/32 = 0.47$

Fisher's Exact Test - $P=0.1347$

Post-Residency Subjects vs. Medical Record Diagnosis

| Decision | Abnormal | Normal | Total |
|-------------|----------|--------|-------|
| Refer | 22 | 58 | 80 |
| No Referral | 20 | 68 | 88 |
| Total | 42 | 126 | 168 |

Table 10: Post-Residency Subjects vs. Medical Record Diagnosis

Sensitivity: $22/42 = 0.52$

Specificity: $68/126 = 0.54$

Fisher's Exact Test - $P=0.4825$

5.2.2. Preliminary Diagnosis of Disorder Populations

Table 11 through Table 14 show results of the preliminary diagnoses for TrendX and the human subjects vs. the Medical Record Diagnosis for the abnormal cases. Each of the groups referred patients fairly consistently across the different disorder sub-populations. Again, one can see that the referral decisions of TrendX were comparable to that of the aggregate group of human subjects, having overall correct referral percentages of 59.2% vs 59.3%. However, the physicians chose the correct preliminary diagnosis more often (36.4% for TrendX vs 56.0% for the human subjects). Table 13 and Table 14 show why. While the Total Correct Dx% and Correct Dx% (of Refs) for TrendX and the Pre-Residency groups are comparable (21.0% vs 23.1% and 36.4% vs 33.3%), the performance of the Post-Residency group was almost twice as good as TrendX (37.9% vs. 21.0% and 68.8% vs. 36.4%).

Disorder Population Referral and Diagnosis Results TrenDx vs. Medical Record

| Disorder Sub Population | # | No Refer | Refer | Correct Ref % | Correct Dx | Correct Dx % | Correct Dx % (of Refs) ³ |
|-------------------------|----|----------|-------|---------------|------------|--------------|-------------------------------------|
| Precocious Pub | 6 | 2 | 4 | 66.7% | 2 | 33.3% | 50.0% |
| GH-Def / Hypothyroid | 11 | 5 | 6 | 54.5% | 1 | 9.1% | 16.7% |
| Cancer / Complex | 8 | 3 | 5 | 62.5% | N/A | N/A | N/A |
| Short Bone / Turner's | 2 | 1 | 1 | 50% | 1 | 50.0% | 100% |
| Total | 27 | 11 | 16 | 59.2% | 4/19 | 21.0% | 36.4% |

Table 11: Disorder Population Referral and Diagnosis Results, TrenDx vs. Medical Record

Disorder Population Referral and Diagnosis Results All Physicians vs. Medical Record

| Disorder Sub Population | # | No Refer | Refer | Correct Ref % | Correct Dx | Correct Dx % | Correct Dx % (of Refs) |
|-------------------------|----|----------|-------|---------------|------------|--------------|------------------------|
| Precocious Puberty | 20 | 9 | 11 | 55.0% | 6 | 30.0% | 54.5% |
| GH-Def / Hypothyroid | 19 | 6 | 13 | 68.4% | 8 | 42.1% | 61.5% |
| Cancer / Complex | 17 | 7 | 10 | 58.8% | N/A | N/A | N/A |
| Short Bone / Turner's | 3 | 2 | 1 | 33.3% | 0 | 0% | 0% |
| Total | 59 | 24 | 35 | 59.3% | 14/42 | 33.3% | 56.0% |

Table 12: Disorder Population Referral and Diagnosis Results,

All Human Subjects vs. Medical Record

³ Correct Dx% of refs does not include Cancer/Complex cases. Neither does Total.

**Disorder Population Referral and Diagnosis Results
Pre-Residency vs. Medical Record**

| Disorder Sub Population | # | No Refer | Refer | Correct Ref % | Correct Dx | Correct Dx % | Correct Dx % (of Refs) |
|-------------------------|----|----------|-------|---------------|------------|--------------|------------------------|
| Precocious Pub | 6 | 2 | 4 | 66.7% | 1 | 16.7% | 25.0% |
| GH-Def / Hypothyroid | 6 | 1 | 5 | 83.3% | 2 | 33.3% | 40.0% |
| Cancer / Complex | 4 | 0 | 4 | 100% | N/A | N/A | N/A |
| Short Bone / Turner's | 1 | 1 | 0 | 0% | 0 | 0% | 0% |
| Total | 17 | 4 | 13 | 76.5% | 3/13 | 23.1% | 33.3% |

*Table 13: Disorder Population Referral and Diagnosis Results,
Pre-Residency vs. Medical Record*

**Disorder Population Diagnosis Results
Post residency vs. Medical Record**

| Disorder Sub Population | # | No Refer | Refer | Correct Ref % | Correct Dx | Correct Dx % | Correct Dx % (of Refs) |
|-------------------------|----|----------|-------|---------------|------------|--------------|------------------------|
| Precocious Puberty | 14 | 7 | 7 | 50.0% | 5 | 35.7% | 71.4% |
| GH-Def / Hypothyroid | 13 | 5 | 8 | 61.5% | 6 | 46.2% | 75.0% |
| Cancer / Complex | 13 | 7 | 6 | 46.2% | N/A | N/A | N/A |
| Short Bone / Turner's | 2 | 1 | 1 | 50% | 0 | 0% | 0% |
| Total | 42 | 20 | 22 | 52.4% | 11/29 | 37.9% | 68.8% |

Table 14: Disorder Population Referral and Diagnosis Results, Post-Residency vs. Medical Record

5.2.3. Summary

The performance of TrenDx and the human subjects at the task of deciding whether to refer patients was comparable, though the specificity of TrenDx was lower. The Pre-Residency subjects had the highest sensitivity, but were only able to choose the correct preliminary diagnosis about as often as TrenDx. The Post-Residency subjects had the lowest sensitivity and highest specificity. Even though they had the lowest sensitivity, they had the highest percent of correct preliminary diagnoses. None of the Chi-Square Tests resulted in statistically significant results.

5.3. Test Subjects vs. Experts

5.3.1. Referral Decision

Table 19 compares the referral decisions of TrenDx to the Expert Gold-Standard. Compared the results shown in Table 7, the sensitivity and specificity are slightly improved (0.61 vs. 0.59 and 0.52 vs 0.47).

TrenDx Decision vs. Expert Gold-Standard

| Decision | Expert Refer | Expert Not Refer | Total |
|-------------|--------------|------------------|-------|
| Refer | 36 | 17 | 53 |
| No Referral | 23 | 19 | 42 |
| Total | 59 | 36 | 95 |

Table 15: TrenDx Decision vs. Expert Gold-Standard

Sensitivity: $36/59 = 0.61$

Specificity: $19/36 = 0.52$

Fisher's Exact Test - $P=0.2080$

Table 16 shows the results of the same analysis performed for the human subjects. Note that their performance is also improved and the specificity values are consistently higher than those of TrenDx. The results of Fisher's Exact Test indicate a statistically significant correlation between the responses of the physicians and those of the experts at $\alpha=0.05$.

Physicians vs. Expert Gold-Standard

| Decision | Expert Refer | Expert Not Refer | Total |
|-------------|--------------|------------------|-------|
| Refer | 91 | 19 | 110 |
| No Referral | 51 | 56 | 107 |
| Total | 142 | 75 | 217 |

Table 16: Physicians vs. Expert Gold-Standard

Sensitivity: $91/142 = 0.64$

Specificity: $56/75 = 0.75$

Fisher's Exact Test - $P < 0.0001$

5.3.2. Referral Decision on Expert and Medical Record Consensus Cases

In this section, we compare TrenDx and the physicians to the set of cases in which the two gold-standards, the experts and the medical record, agreed. Table 17 shows the results for TrenDx, and Table 18 shows the results for the physicians. The middle columns in both tables represent the cases in which the expert agreed with the medical record diagnosis, either by referring an abnormal case or not referring a normal case.

For TrenDx, the consensus sensitivity (sensitivity over patients in whom both gold-standards agreed should be referred) was higher than the sensitivity vs. the expert (0.68 vs. 0.61, Table 17). The specificity was lower than the specificity compared solely to the expert (0.50 vs 0.52) but still higher than the specificity compared to the medical record diagnosis (0.50 vs. 0.47 in Table 7). Again, the P-values were not statistically significant.

TrenDx Decision vs. Expert and Medical Record Consensus

| Decision | Expert Refer | | Expert Not Refer | | Total |
|-------------|--------------|---------------|------------------|--------|-------|
| | Normal | Abnorm | Normal | Abnorm | |
| Refer | 22 | 13 | 14 | 3 | 53 |
| No Referral | 18 | 6 | 14 | 5 | 42 |
| Total | 40 | 19 | 28 | 8 | 95 |

Table 17: TrenDx Decision vs. Expert and Medical Record Consensus

Consensus Sensitivity: $13/19 = 0.68^4$

Consensus Specificity: $14/28 = 0.50$

Fisher's Exact Test - $P=0.2438$

The decisions of the physicians correlate much better to the consensus decisions (Table 18) than either the expert or medical record decisions individually.

Physicians vs. Expert and Medical Record Consensus

| Decision | Expert Refer | | Expert Not Refer | | Total |
|-------------|--------------|---------------|------------------|--------|-------|
| | Normal | Abnorm | Normal | Abnorm | |
| Refer | 59 | 32 | 16 | 3 | 110 |
| No Referral | 40 | 11 | 43 | 13 | 107 |
| Total | 99 | 43 | 59 | 16 | 217 |

Table 18: Physicians vs. Expert and Medical Record Consensus

Consensus Sensitivity: $32/43 = 0.74^5$

Consensus Specificity: $43/59 = 0.73$

Fisher's Exact Test - $P < 0.0001$

5.3.3. Preliminary Diagnosis of All Cases

The number of times that the preliminary diagnosis of the test subjects matched to those of the experts was counted and is shown in Table 19.

⁴ Sensitivity and Specificity calculated out of the two middle columns representing the consensus of Expert and Medical Record

⁵ Sensitivity and Specificity calculated out of the two middle columns representing the consensus of Expert and Medical Record

Test Subject Preliminary Dx Matches to Expert Dx

| Group | # Decisions | # Match | % Match |
|--------------------|-------------|---------|---------|
| TrenDx | 95 | 39 | 41.0% |
| All Human Subjects | 217 | 123 | 56.7% |

Table 19: Test Subject Preliminary Diagnosis Matches to Expert Diagnosis

5.3.4. Summary

In summary, the test subjects performed better in almost all areas when compared to the expert gold-standard instead of the medical record. In addition, the performance of the human subjects was even better over the cases in which the expert and the medical record agreed. The performance of TrenDx for those cases was only slightly improved. Not surprisingly, the referral decisions of the human subjects matched those of the experts more often than the decisions of TrenDx matched the experts.

5.4. TrenDx and Human Subject Scores

The responses of all the subjects were scored according to the algorithm described in section 3.3. The entire set of scores for all the participants is listed in **Appendix A - Patient and Subject Result Tables**. Table 20 shows the average score received by each subject group - TrenDx, all human participants, the Pre-Residency group, and the Post-Residency group. The average score of TrenDx was lower than the average scores of any of the participant groups. The two Pre-Residency and Post-Residency groups performed comparably. The table also shows the 95% confidence interval calculations and the results of the Wilcoxon Signed-Rank Test comparing the human subject scores to the scores received by TrenDx. The differences are statistically significant for the entire group of participants and for the Post-Residency participants at $\alpha=0.05$, indicating a high probability that TrenDx does not score as well as the other groups at this task. Remember that the score is based on the expert gold-standard and we have already shown that there is a statistically significant correlation between the decisions of the expert gold-standard and the human subjects.

Test Subject Scores

| Test Subject Group | # Decisions | Avg±S.D. | 95% C.I. | t-test P Value |
|--------------------|-------------|----------|----------|---------------------|
| TrenDx | 95 | 4.7±0.8 | 3.7-5.6 | N/A |
| All Humans | 217 | 5.4±1.2 | 4.9-5.9 | 0.0033 ⁶ |
| Pre-Residency | 49 | 5.5±0.6 | 4.8-6.2 | 0.0625 |
| Post-Residency | 168 | 5.4±1.3 | 4.7-6.0 | 0.0135 ⁷ |

Table 20: Test Subject Scores

The scores of TrenDx and the human subjects over the different disorder sub-populations are presented in Table 21. The two groups performed comparably on the Precocious Puberty cases, but the physicians scored considerably better on the GH-Def/Hypothyroid cases and the Complex / Cancer cases. TrenDx scored better on the Short Bone/Turner's cases, but there were very few of those cases. At the bottom of the table, the weighted average of the scores is shown. Note that the weighted average score of TrenDx on these abnormal cases is the same as the average score of TrenDx over all cases (Table 20). This is in contrast to the physicians; their weighted average score on the abnormal cases is much higher than their score over all cases. The weighted average score of the physicians on the normal population is 5.18.

Average Score by Disorder Sub-Population

| Disorder | # Pats | TrenDx Avg Score | # Dec | Physician Avg Score |
|----------------------|--------|------------------|-------|---------------------|
| Precocious Puberty | 6 | 5.8 | 20 | 5.6 |
| GH-Def / Hypothyroid | 11 | 4.4 | 19 | 6.0 |
| Complex / Cancer | 8 | 4.2 | 17 | 6.9 |
| SB/Turner's | 2 | 5 | 3 | 3.3 |
| Total / Weighted Avg | 27 | 4.7 | 59 | 6.0 |

Table 21: Average Score by Disorder Sub-Population

5.4.1. Summary

Using the scoring algorithm, the human subjects generally performed better than TrenDx, earning higher average scores over all the patient cases and when the patient cases

⁶ Statistically Significant

⁷ Statistically Significant

were divided into normal/abnormal cases. The difference in scoring was statistically significant at $\alpha=0.05$.

5.5. Multiple Comparisons of Test Subjects and Gold Standards

Of the 95 patient cases in this trial, 59 of them had more than one human subject/reviewer. Out of those 59 patient cases, there was unanimous consensus among 29 of the human subjects in terms of the referral decision, with 18 referrals and 11 patients not referred. For each of those 29 cases, we looked at the combination of the medical record diagnosis, the expert's decision, TrenDx's decision, and the consensus decision of the physicians. Among the 29 cases, there were 4 cases in which every group did not refer the patient. There were 4 cases in which every group did refer the patient.

Then we looked at cases in which there were singular decisions (one group made one decision and all other groups made the opposite decision). There were 9 cases in which the medical record stated that the patient was normal but all of the other groups referred the patient. There were three cases in which the experts had a singular decision; they were all decisions to refer. There were 2 cases in which the decision of TrenDx differed from those of all the other groups. In one, TrenDx referred and in the other, TrenDx did not refer. The human subject group had no singular decisions. There were 7 remaining decisions in which two of the groups referred and two didn't. These results are summarized in Table 22.

It is interesting to look more closely at the 9 cases in which the medical record categorized the patient as normal but every other group felt that the patient should be referred. Looking at each of those cases individually, it is clear that a case can be made for the referral decision in each case. In 6 of those cases, the patient's height dropped a significant number of percentiles, with the smallest drop being 25 percentile points, and most of the cases falling to well below the fifth percentile. Children that short are abnormal by definition, being well below 2 standard deviations from the mean. In the three other cases, one was an infant with a bone age that was almost a year greater than the child's chronological age, one was a child with a bone age that was advanced by three years, and the third was a child whose height went from the 95th percentile to well above the 95th percentile. These cases suggest that there is a difference between being "normal" and not having a medical condition or disorder. If a child has no medical condition, but his height is three standard deviations away from the mean, is he normal?

Consensus and Singular Referral Decisions

| Description | Number |
|---|--------|
| All groups agree - Refer | 4 |
| All groups agree - No referral | 4 |
| Normal (MR), all others referred | 9 |
| Expert refer, no one else referred | 3 |
| TrenDx refer, no one else referred | 1 |
| TrenDx no referral, all others referred | 1 |
| Split decision | 7 |
| Human subject singular decisions | 0 |

Table 22: Consensus and Singular Referral Decisions

There are other interesting things to note. For example, there were only 2 cases in which the medical record indicated that a referral was necessary, but none of the other groups referred the patient. Only 1 human subject reviewed those patient cases. There were 3 cases in which the patient was abnormal and only TrenDx triggered a referral (including experts). And there were 3 cases in which the patient was abnormal and only one of the human subjects referred the case. There were no cases in which the patient was abnormal and the expert was the only one to refer the case.

5.5.1. Summary

Looking purely at the number of singular decisions for the 29 cases that had consensus among their human subject reviewers, there were 9 for the medical record, 3 for the experts, 2 for TrenDx, and none for the human subjects.

5.6. Variations in Threshold Triggering

As described in Section 4.3.4, the threshold triggering values were obtained by processing twenty 'normal' cases and using the lowest error-score of the trend-templates for Normal Growth, Constitutional Delay of Puberty, and Early Puberty. The threshold for single point triggering was set at 0.35, which was 2 standard deviations from the mean.

The threshold for consecutive triggering was set at 0.31, approximately 1.5 standard deviations from the mean.

This section describes the results obtained from raising and lowering the threshold triggering values by one half of the standard deviation, and the effects of ignoring certain data points.

5.6.1. Raising Threshold Triggering Values

In the first test, both thresholds were raised by one half a standard deviation. The single-point triggering threshold became 0.39 and the consecutive-point triggering threshold became 0.35. Table 23 shows the results of raising the triggering thresholds. Recall that Trendx referred 50 of the 95 cases. Raising the triggering thresholds prevented 7 of the 50 cases from triggering a referral. Of those 7 patients, 6 were normal and 1 was abnormal.

Results of Raising Triggering Thresholds

| Decision | Abnormal | Normal |
|-------------|----------|--------|
| Refer | 15 | 28 |
| No Referral | 11 | 38 |

Table 23: Results of Raising Triggering Thresholds

Sensitivity: 0.56⁸

Specificity: 0.56

Chi-Square Test - P=0.2468

5.6.2. Lowering Threshold Triggering Values

Then, the effects of lowering the triggering thresholds by 1/2 a standard deviation was examined. The single-point threshold was lowered to 0.31 and the consecutive-point threshold was lowered to 0.27. These values represent 1.5 and 1 standard deviation from the mean error-value, respectively. Using the new, lowered triggering thresholds, 7 new referrals occurred. Of these, 2 were abnormal patients that deserved to be referred and 5 were normal patients. The results are shown in Table 24.

⁸ Sensitivity and Specificity calculated out of 27 abnormal and 68 normals.

Results of Lowering Triggering Thresholds

| Decision | Abnormal | Normal |
|-------------|----------|--------|
| Refer | 18 | 39 |
| No Referral | 8 | 27 |

Table 24: Results of Lowering Triggering Thresholds

Sensitivity: 0.67

Specificity: 0.40

Chi-Square Test - P= 0.997

5.6.3. Ignoring Particular Error-Scores

We examined the results of “ignoring” the referral if it was made on the first, non-trivially scoring data point. In that case, there were 7 fewer referrals. However, 2 of those 7 were abnormal patients who should have been referred. Table 25 lists the resulting sensitivity and specificity.

Results of Ignoring First Non-Trivial Point

| Decision | Abnormal Patients | Normal Patients |
|-------------|-------------------|-----------------|
| Refer | 14 | 29 |
| No Referral | 12 | 37 |

Table 25: Results of Ignoring First Non-Trivial Point

Sensitivity: $14/27 = 0.52$

Specificity: $37/68 = 0.54$

Fisher’s Exact Test - P=0.4877

Results of Ignoring Infant Scores

| Decision | Abnormal Patients | Normal Patients |
|-------------|-------------------|-----------------|
| Refer | 15 | 28 |
| No Referral | 11 | 38 |

Table 26: Results of Ignoring Infant Scores

Sensitivity: $15/27 = 0.56$

Specificity: $38/68 = 0.56$

Fisher’s Exact Test - P=0.2468

In addition, it was noted that the error-scores for infants were somewhat higher than for adults. This suggested that the trend-templates did not model infancy well enough. We considered the results of ignoring infant scores and only looking at childhood data points (> age 3) for those patients that had childhood data. Out of the patients whose error-scores triggered a referral, ignoring the infant data point error-scores prevented 7 of them from being referred. Of those 7, only 1 was abnormal and deserved to be referred. The consequences of ignoring infant data points (< age 3) are presented in Table 26.