TIME SERIES ANALYSIS AS INPUT FOR PREDICTIVE MODELING: PREDICTING CARDIAC ARREST IN A PEDIATRIC INTENSIVE CARE UNIT

Curtis Kennedy
*University of Texas Health Science Center at Houston*

Follow this and additional works at: https://digitalcommons.library.tmc.edu/uthshis_dissertations

Part of the Medicine and Health Sciences Commons

Recommended Citation

This is brought to you for free and open access by the School of Biomedical Informatics at DigitalCommons@TMC. It has been accepted for inclusion in UT SBMI Dissertations (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact nha.huyn@library.tmc.edu.
TIME SERIES ANALYSIS AS INPUT FOR PREDICTIVE MODELING:
PREDICTING CARDIAC ARREST
IN A PEDIATRIC INTENSIVE CARE UNIT

By

Curtis Kennedy, MD, MS

October 5, 2010

APPROVED:

James P. Turley, PhD, RN
Jack W. Smith, MD, PhD
Noriaki Aoki, MD, PhD
M. Michele Mariscalco, MD
TIME SERIES ANALYSIS AS INPUT FOR PREDICTIVE MODELING:

PREDICTING CARDIAC ARREST

IN A PEDIATRIC INTENSIVE CARE UNIT

A

DISSERTATION

Presented to the Faculty of
The University of Texas
School of Biomedical Informatics
at Houston
in Partial Fulfillment
of the Requirements

for the Degree of
Doctor of Philosophy

by

Curtis Edward Kennedy

Committee Members:

James P. Turley, Ph.D., R.N.,
Jack W. Smith, M.D., Ph.D.,
Noriaki Aoki, M.D., Ph.D.,
M. Michele Mariscalco, M.D.

1 The University of Texas School of Biomedical Informatics at Houston
2 Baylor College of Medicine, Department of Pediatrics, Section of Critical Care Medicine
Copyright
by
Curtis Kennedy
2010
DEDICATION

To those who have inspired, believed in, and supported my efforts
…to craft a dream from a collection of ideas
…and to help a dream come true,
I dedicate this work.

To the children who have suffered,
my deepest respect for your sacrifice.
You are my touchstone.
# TABLE OF CONTENTS

Dedication ................................................................................................................4
Table of Contents .....................................................................................................5
Acknowledgements ..................................................................................................6
Introduction ..............................................................................................................8
Manuscripts .............................................................................................................15
  Time-Series Analysis as Input for Clinical Predictive Modeling:  
  Modeling Cardiac Arrest in a Pediatric ICU.......................................................16
  Building Clinical Prediction Models Using Time Series Data:  
  Modeling Cardiac Arrest in a Pediatric ICU.....................................................60
  Using Time Series Analysis to Predict Cardiac Arrest  
  in a Pediatric Intensive Care Unit .................................................................92
Summary ..............................................................................................................124
ACKNOWLEDGEMENTS

I entered subspecialty training in pediatric critical care medicine as a fellow in July of 2000. At the time I knew I liked computers and math, and I desperately wanted to align these likes with the research requirements of my fellowship. Informatics was a pretty new concept at the time, and it took quite a bit of effort to piece together something that passed as research. It was only through the actions of Dr. Larry Jefferson, my former section chief, Dr. Robert (Bob) Beck, my initial mentor, Dr. Noriaki Aoki, my initial guide in data mining and ongoing mentor, and Dr. Jack Smith, my mentor after Dr. Beck took another job, that an infrastructure to support a research career in informatics was made available to me. These people saw something that I could not see at the time. They believed in me and they have taken chances on multiple occasions to get me where I am today. For their help, I am eternally grateful.

Dr. Michele Mariscalco was my fellowship director and is my ongoing mentor. She is responsible for helping me translate “I like computers and math” into hypothesis driven research. Her patience in reviewing my countless research proposals for my fellowship, and her vigilant critique of all my subsequent endeavors has been pivotal in my growth as an academic physician.

The faculty at the School of Health Information Sciences (now the School of Biomedical Informatics) has been exceedingly supportive over the past decade of my informatics training. They have provided me with an excellent background, both theoretical and practical, in a diverse number of topics germane to informatics. Their encouragement and infective enthusiasm has been the staying force that has kept me going all this time. I thank the entire faculty for this. In particular, Dr. Smith has been instrumental in helping me anticipate and overcome barriers that I
have encountered along the way. Most of my successes can be traced back to his influence at some level. Also, I credit Dr. James (Jim) Turley, my PhD committee chair, with helping me articulate the theoretical elements of this work. Before his involvement I was very goal and task oriented. He helped me recognize my accomplishments as something more than a task.

On a more global level, this work has been supported by Baylor College of Medicine, which has provided me with the protected time and human resources required to establish a foundation for this research. It has also been supported by Texas Children’s Hospital, who has provided hardware support and the human resources required for assimilating the raw data. The National Library of Medicine Medical Informatics Resource Training Grant (5T15 LM07093) provided additional resources and training to help make this work possible, as did a K-22 career development award (1K22 LM008389). All of these resources have provided me with a foundation on which to build a research career, and I am indebted to them all.

Finally, I would be remiss if I failed to include the most important members of the team. Trlica, my loving wife: I would never have been able to do this without your unremitting support. I cannot thank you enough. My love for you is superlative. Celeste and Jackson, my adorable children: Thank you for being so good when Mommy had to solo cover you two. I am a very lucky Daddy to have you kiddos. Don, Jeanne, and Kelly, my dad, mom, and brother: You always provide love, support, and encouragement, especially when I need it the most. Thank you for all you do and for all you have done.
INTRODUCTION

It took many years for me to narrow my research focus to predicting cardiac arrest in pediatric intensive care units. Every new idea and epiphany I came up with along the way generated advice to refine the question and to narrow the focus to something more specific, more manageable. So, I continued learning about the ideas I was pitching, and I continued narrowing my focus. Data mining evolved to knowledge discovery, then to predictive modeling. Outcomes evolved to mortality, then to cardiac arrest. Cardiac arrest seemed basic enough, but even that topic had to have additional constraints added: target those attributable to progressive shock, and leave for later those caused by other factors. The quote by Nobel laureate Konrad Lorenz kept coming to mind: “Every man gets a narrower and narrower field of knowledge in which he must be an expert in order to compete with other people. The specialist knows more and more about less and less and finally knows everything about nothing.”

After many unsuccessful attempts and revisions, I eventually crafted a nice, tight, manageable task. I set out to complete it. I was grateful to my advisors for their insistence to focus my efforts – it made my task very clear. It helped me sell the idea I was pitching. I was happy, and a bit proud of what I was doing. Then critique started coming in. Apparently, I had completed a very focused task but had failed to contribute anything meaningful to the scientific field. After all that hard work guided by Lorenz inspired blinders, now I was being criticized for failing to meet Schrödinger inspired criteria. (Schrödinger was also a Nobel laureate who is best known for his cat analogy in quantum theory, but was also a prominent philosopher. The quote I refer to here is: “…the isolated knowledge obtained by a group of specialists in a narrow field has in itself no value whatsoever, but only in its synthesis with all the rest of knowledge and only inasmuch as it really contributes in this synthesis toward answering the demand…” This was a
very low point in my academic career. I was derailed for over a year. I held religiously to the belief that my creation was new and potentially useful and therefore was a contribution to the field. Still, the people that were providing the critique were smart people, and I could find no fundamental fault in what they were telling me.

Acknowledging that there are a wide range of religious beliefs and non-beliefs, and recognizing that religion’s presence in scientific work is rarely encouraged, I feel compelled to credit God with my emergence from the realm of despair. Others may feel more comfortable attributing it to natural processes or luck. At any rate, several months after my setback one of my advisors, Dr. Hualou Liang, ended up pursuing a job in another city. He left my committee vacant one seat, and I sought to fill the vacancy with someone sympathetic to my plight. Fortunately, I found no one willing to champion my cause as I saw it. However, I did engage in several important conversations with Dr. Jim Turley, who ultimately filled the vacancy and has since served as my committee chair. After Dr. Turley listened to my story and my arguments, he asked me to set my self evaluations aside, to step back, and to break apart what I had done into piece by piece units. It seemed at the time to be a bit of an obscure exercise, but it was clear he was asking me to do it for a reason. After several iterations, I finally decomposed my work enough for him to smile. At that point, he asked me to tell him my story again, this time using the template that I had prepared. When I was done, he said “That’s it!” At that moment, I thought one of us must be insane. Then he instructed me to retell my story, except instead of using cardiac arrest in a pediatric intensive care unit, to use another scenario. (Admittedly, I don’t recall the specific scenario… perhaps respiratory failure in an adult intensive care unit or renal failure in an acute care setting.) The epiphany hit me like a ton of bricks. I had been so focused on the task of predicting cardiac arrest and its specifics that I had failed to recognize the broader
implications. It was at this point that I realized my research fundamentally dealt with a theoretical strategy for incorporating time series data into a predictive modeling task. Predicting cardiac arrest in a pediatric intensive care unit served as the illustrative case, not the sole objective.

Armed with this paradoxical concept of being able to broadly apply what I had worked so hard to narrow, limit and focus, I have prepared three manuscripts for publication. They cover the theoretical background behind using time series data elements in a predictive modeling task, the method of applying the theory using cardiac arrest prediction in a pediatric intensive care setting as an example, and the results achieved by carrying out the described methods. In addition to sharing what I have learned with the scientific community, these manuscripts are intended to fulfill the dissertation requirement for a doctorate of philosophy degree in health informatics.

The first manuscript, entitled “Time-Series Analysis as Input for Clinical Predictive Modeling: Modeling Cardiac Arrest in a Pediatric ICU” lays out the theoretical background for the project. There are several core concepts presented in this paper. First, traditional multivariate models (where each variable is represented by only one value) provide single point-in-time snapshots of patient status: they are incapable of characterizing deterioration. Since deterioration is consistently identified as a precursor to cardiac arrests, we maintain that the traditional multivariate paradigm is insufficient for predicting arrests. We identify time series analysis as a method capable of characterizing deterioration in an objective, mathematical fashion, and describe how to build a general foundation for predictive modeling using time series analysis results as latent variables. Building a solid foundation for any given modeling task involves addressing a number of issues during the design phase. These include selecting the proper
candidate features on which to base the model, and selecting the most appropriate tool to measure them. We also identified several unique design issues that are introduced when time series data elements are added to the set of candidate features. One such issue is in defining the duration and resolution of time series elements required to sufficiently characterize the time series phenomena being considered as candidate features for the predictive model. Once the duration and resolution are established, there must also be explicit mathematical or statistical operations that produce the time series analysis result to be used as a latent candidate feature. In synthesizing the comprehensive framework for building a predictive model based on time series data elements, we identified at least four classes of data that can be used in the model design. The first two classes are shared with traditional multivariate models: multivariate data and clinical latent features. Multivariate data is represented by the standard one value per variable paradigm and is widely employed in a host of clinical models and tools. These are often represented by a number present in a given cell of a table. Clinical latent features derived, rather than directly measured, data elements that more accurately represent a particular clinical phenomenon than any of the directly measured data elements in isolation. The second two classes are unique to the time series data elements. The first of these is the raw data elements. These are represented by multiple values per variable, and constitute the measured observations that are typically available to end users when they review time series data. These are often represented as dots on a graph. The final class of data results from performing time series analysis. This class of data represents the fundamental concept on which our hypothesis is based. The specific statistical or mathematical operations are up to the modeler to determine, but we generally recommend that a variety of analyses be performed in order to maximize the likelihood
that a representation of the time series data elements is produced that is able to distinguish between two or more classes of outcomes.

The second manuscript, entitled “Building Clinical Prediction Models Using Time Series Data: Modeling Cardiac Arrest in a Pediatric ICU” provides a detailed description, start to finish, of the methods required to prepare the data, build, and validate a predictive model that uses the time series data elements determined in the first paper. One of the fundamental tenets of the second paper is that manual implementations of time series based models are unfeasible due to the relatively large number of data elements and the complexity of preprocessing that must occur before data can be presented to the model. Each of the seventeen steps is analyzed from the perspective of how it may be automated, when necessary. We identify the general objectives and available strategies of each of the steps, and we present our rationale for choosing a specific strategy for each step in the case of predicting cardiac arrest in a pediatric intensive care unit. Another issue brought to light by the second paper is that the individual steps required to use time series data for predictive modeling are more numerous and more complex than those used for modeling with traditional multivariate data. Even after complexities attributable to the design phase (addressed in our first paper) have been accounted for, the management and manipulation of the time series elements (the preprocessing steps in particular) are issues that are not present in a traditional multivariate modeling paradigm. In our methods, we present the issues that arise from the time series data elements: defining a reference time; imputing and reducing time series data in order to conform to a predefined structure that was specified during the design phase; and normalizing variable families rather than individual variable instances.

The final manuscript, entitled: “Using Time-Series Analysis to Predict Cardiac Arrest in a Pediatric Intensive Care Unit” presents the results that were obtained by applying the
theoretical construct and its associated methods (detailed in the first two papers) to the case of cardiac arrest prediction in a pediatric intensive care unit. Our results showed that utilizing the trend analysis from the time series data elements reduced the number of classification errors by 73%. The area under the Receiver Operating Characteristic curve increased from a baseline of 87% to 98% by including the trend analysis. In addition to the performance measures, we were also able to demonstrate that adding raw time series data elements without their associated trend analyses improved classification accuracy as compared to the baseline multivariate model, but diminished classification accuracy as compared to when just the trend analysis features were added (ie, without adding the raw time series data elements). We believe this phenomenon was largely attributable to overfitting, which is known to increase as the ratio of candidate features to class examples rises. Furthermore, although we employed several feature reduction strategies to counteract the overfitting problem, they failed to improve the performance beyond that which was achieved by exclusion of the raw time series elements. Finally, our data demonstrated that pulse oximetry and systolic blood pressure readings tend to start diminishing about 10-20 minutes before an arrest, whereas heart rates tend to diminish rapidly less than 5 minutes before an arrest.

In conclusion, the three manuscripts that are attached and have been submitted for publication constitute a dissertation detailing the work that has been performed in leveraging time series data to improve the performance of clinical prediction models by characterizing important features, such as trends over time, that are not addressed by traditional multivariate modeling paradigms. The methods and theoretical constructs that are presented are intended to serve a broad class of clinical prediction needs where historical data influences the interpretation of any given condition. We have demonstrated its effectiveness in a case of predicting cardiac
arrest in a pediatric intensive care unit, but have only evaluated a relatively small amount of potentially useful information contained in time series data. We believe that other time series analyses that have been presented from a theoretical perspective, but have not yet been demonstrated by example, hold great promise for serving as the basis for a new breed of clinical prediction tools that can continuously assess time series data and provide clinicians with useful information that can help them provide better, safer care to the patients they serve.

Curtis Kennedy, M.D., M.S.
October 5, 2010
MANUSCRIPTS
Time-Series Analysis as Input for Clinical Predictive Modeling: Modeling Cardiac Arrest in a Pediatric ICU

Curtis Edward Kennedy, M.D., M.S.¹
James P. Turley, Ph.D., R.N.²

¹Baylor College of Medicine, Department of Pediatrics, Section of Critical Care Medicine
One Baylor Plaza, Houston, TX 77030
²The University of Texas Health Science Center at Houston – School of Biomedical Informatics,
7000 Fannin Suite 600, Houston, TX 77030

Corresponding author
Curtis Kennedy, MD, MS
cokenned@texaschildrenshospital.org
Baylor College of Medicine
Department of Pediatrics
Critical Care Section
One Baylor Plaza
Houston, TX 77030
832-826-6230
Abstract

Background

Thousands of children experience cardiac arrest events every year in pediatric intensive care units. Most of these children die. Cardiac arrest prediction tools are used as part of medical emergency team evaluations to identify patients in standard hospital beds that are at high risk for cardiac arrest. There are no models to predict cardiac arrest in pediatric intensive care units though, where the risk of an arrest is 10 times higher than for acute care units. Current tools are based on a multivariate approach that does not characterize deterioration, which often precedes cardiac arrests. Deterioration requires a time series approach in order to characterize.

The purpose of this study is to propose a method that will allow for time series data to be used in clinical prediction models. Successful implementation of these methods has the potential to bring arrest prediction to the pediatric intensive care environment, possibly allowing for interventions that can save lives and prevent disabilities.

Methods

We reviewed prediction models from nonclinical domains that employ time series data, and identified the steps that are necessary for building predictive models using time series clinical data. We illustrate the method by applying it to the specific case of building a predictive model for cardiac arrest in a pediatric intensive care unit.

Results

Time course analysis studies from genomic analysis provided a modeling template that was compatible with the steps required to generate a model from clinical time series data. The steps
include: 1) selecting candidate variables; 2) specifying measurement parameters; 3) defining data format; 4) defining time window duration and resolution; 5) calculating latent variables for candidate variables not directly measured; 6) calculating time series features as latent variables; and 7) creating data subsets to measure model performance effects attributable to various classes of candidate variables.

Conclusions

We have proposed a seven step process that results in data sets that contain time series features and are suitable for predictive modeling by a number of methods. We illustrated the process through an example of cardiac arrest prediction in a pediatric intensive care setting.

Background

Roughly 1-6% of children being cared for in an ICU will experience a cardiac arrest while in the ICU.(1, 2) Many of these arrests occur because their vital signs deteriorate to the point where they enter a state of progressive shock.(3-5) These arrests happen despite the fact that they are being continuously monitored by EKG, pulse oximetry, and frequent blood pressure measurements. There are tools to help identify children in acute care units that have deteriorated to the point where they need to transfer to an ICU(6-8), but once in the ICU there are no tools to identify which children are likely to arrest. Having already met the criteria for “high risk” and been transferred to an ICU, it is incumbent on bedside caregivers to detect further deteriorations, distinguish the life threatening deteriorations from routine variability, and provide preventative treatment. It is an imperfect science that is operator dependent and no single threshold works from patient to patient or even for a single patient from one point in time to another. The scoring tools that work in acute care units to distinguish healthy from sick use single measurements and
set thresholds in order to measure level of illness. They are multivariate in nature: one value per variable. They do not characterize “deterioration” per se; they only characterize the end effect. The end effect of deterioration in an ICU is cardiac arrest. Tools that are able to characterize “deterioration” are needed, but since deterioration is a time dependent phenomenon, the tools need to accommodate time series data: multiple values per variable. The goal of this study is to develop a framework for building prediction models that use time series data and can serve as the foundation for tools that can characterize deterioration, with the hope of someday being able answer questions like “Who is most likely to arrest?” in an ICU environment.

Clinicians evaluate data to gain an understanding of its implications so they can provide care and education to their patients. If the clinician cannot gain understanding from the data, additional data is typically sought. The additional data often comes in the form of additional testing, but it can comes in the form of tools that help the clinician understand existing data. In their simplest form, tools can be as basic as a growth chart that informs the clinician about the patient’s height relative to the population (as a percentage) rather than just the raw number of centimeters. More sophisticated tools can integrate dozens of data elements and inform the clinician about measures ranging from risk of death during a hospitalization to the need to transfer to an intensive care unit(9-12).

There are over 13,000 tools available to help clinicians understand the data they presented(13). Almost all of these tools have been designed so that they can be manually used. A tool’s success in adoption typically depends on a balance between how easy it is to use and what the informational content of the tool is(14, 15) so tools that are built to be manually used are constrained to a relatively small number of variables in order to achieve adequate simplicity. As a result, input variables have typically been restricted to a multivariate data paradigm where each
variable is represented by a single value. A consequence of this strategy is that useful trend information cannot be incorporated into a model unless it is explicitly encoded as a variable. Of course doing this would add complexity to the task, so it is therefore rarely done.

As healthcare is transitioning from manual processes to electronic ones, it is becoming increasingly easy to automate the processes of data collection and analysis. In an automated system, there is no longer a need to remain constrained to a multivariate data paradigm in order to achieve simplicity at the user level. Clinical studies using time series analysis has been undertaken in a number of settings (16-19), but thus far has been relatively limited in scope, tending to focus on interpretation of a single analytic method rather than incorporating multiple analytic methods into a more robust modeling paradigm.

The purpose of this article is to describe a method for developing clinical prediction models based on time-series data elements. The model development process that we are presenting is novel to clinical medicine, but the individual steps comprising the process are not. Our intention is to provide not only the description of the method, but the theoretical basis of the steps involved. We are demonstrating the application of this process in an example of cardiac arrest prediction in pediatric intensive care patients. It is our hope that we describe the steps of the process and their theoretical basis clearly enough that the methodology can be extended to other domains where predictions based on time-series data is needed.

**Introduction**

In order to ensure that the concepts in this article can be understood by clinician and nonclinician alike, we will provide four brief overviews of the core concepts that form the foundation of this article. First, we will describe how the growth of data has impacted medicine
and some of the strategies that have evolved to manage this growth. Second, we will review a few relevant concepts that relate to statistical analysis and modeling, with special focus on multivariate versus time series data paradigms. Third, we will specifically discuss clinical prediction models: their utilities, their limitations, and considerations for improvements. Finally, we will review the rationale behind selecting inpatient cardiac arrest as the example to illustrate the process, and we will provide a brief overview of the physiologic principles that serve as the theoretical basis for our prediction model.

**Data in Medicine:** Medical care has existed since long before diseases were understood at a scientific level. Early medical care was characterized more by art and religion than by science as we know it today. Diagnosis was not possible until advanced stages when symptoms were severe. Therapies were crude and any ‘efficacy’ they had was often due to treating symptoms rather than curing the disease. If someone fell ill, being able to predict whether or not they would survive, and for how long, was pretty much a guess although it may have been informed by a handful of variables such as the nature of the disease, their physical appearance, their mental state, and the like. Over the centuries, data obtained from careful observations have provided us with information and knowledge and has transformed the field of medicine from one of art and religion to one of science. Diagnoses can be made long before symptoms ever manifest. Although many therapies are still imperfect, we’ve moved beyond treating just the symptoms and can now usually treat their underlying cause as well. If someone falls ill today, being able to predict whether or not they survive, and for how long is still a bit of a guess, but it is now better informed: our fundamental knowledge is more complete, patient data is more abundant, and the results of multitudes of studies have all worked together to result in better information that improves our predictive accuracy and narrows the windows of uncertainty.
While medicine’s transition from art to science has improved outcomes, there is a side effect of the data: information overload\(^\text{21-24}\). Currently, the amount of data in the medical field is so extensive and is growing so fast that is impossible for any single person to utilize it all effectively. In order to utilize data, it must be interpreted in context (transforming it into information) and evaluated by the user\(^\text{25}\). This process requires substantial cognitive resources and is time consuming. In an attempt to address this problem, at least two strategies have been employed: specialization and computerized support\(^\text{21, 26}\). Specialization allows clinicians to focus their efforts on a narrow field where they become expert in a relatively small group of related diseases. In doing so, they reduce their educational burden to a point where they can “afford” the cost of training and staying current in their specialty. Most physicians have a basic understanding of disease categories, but to be a specialist in more than one category is rare. The strategy of specialization is inherently limited by the fact that as the body of data continues to grow, eventually the capacity to master even a group of related diseases will be overwhelmed and further specialization will be required. This is already being evidenced by the fact that some specialties have even reached the point where subspecialization is required in order to stay abreast of the latest trends\(^\text{27}\). For example, a cardiologist is a specialist who is expected to know how to diagnose and treat diseases of the heart, but is not expected to manage liver cirrhosis. An electrophysiologist is a cardiology subspecialist that is expected to know a great deal more about how to diagnose and treat electrical disturbances of the heart than a general cardiologist, but is not expected to manage chronic congestive heart failure.

Since there is a fundamental limitation to specialization as a means to cope with excessive amounts of data or information, a more robust solution to the problem is needed. Ideal properties of the new solution should include: scalability\(^\text{28}\) (it can continue to grow indefinitely),
flexibility (it can be used for a number of purposes), explicit and accurate (it relies on objective parameters), and automaticity (it functions independent of frequent supervision). Computer technology possesses these characteristics, and the field of informatics has been born out of effort to utilize computer based solutions to automate the transformation of data to information in the healthcare setting. These solutions come in many forms, ranging from aggregating knowledge available on a given disease to informing clinicians when tests or treatments violate parameters deemed to be unsafe. One of the fundamental goals of this article is to describe a method that can be automated as a computer based solution to help inform clinicians of a patient’s risk of cardiac arrest using trend information that would otherwise require manual interpretation. Since clinicians cannot continuously check the risk of cardiac arrest for all patients they are caring for, we are attempting to leverage information from data that would otherwise be left unanalyzed in the current “intermittent check” paradigm.

**Statistical Analysis and Modeling:** Of course, medicine is not the only field where data has become so abundant that it is impossible to understand it all. Compared to fields such as physics and astronomy, medicine is in a relative state of adolescence. When presented with an abundance of data, the first priority is to understand what the data represent. This process of gaining an understanding is based on statistical analysis. Depending on the information needs, data can be analyzed in a number of ways to provide a range of understandings. For instance, a univariate analysis of “heart rate” provides an understanding of what the most common heart rate is, the range of heart rates, and how the range is distributed. A multivariate analysis that includes “heart rate” as a variable can provide an understanding of how heart rate relates to temperature or blood pressure. A time series analysis of “heart rate” can provide an understanding of how the heart rate changes different times of the day. The statistical methods
for analyzing the data differ fundamentally for time series data since a single variable is represented by multiple values that vary depending on the time they represent. Univariate and multivariate statistics, on the other hand, rely on a single value per variable for each case. Also, time series data elements are assumed to correlate to adjacent data elements(40), whereas this type of correlation can interfere with univariate and multivariate analysis(41, 42).

Whereas univariate and multivariate data analysis informs the user of the distribution of a variable across a population and how the variable relates to other variables, time series analysis informs the user of how a variable relates to itself. In particular, time series analysis provides two types of information about a variable of interest: trends and seasonality(43). The distinction between the two is that univariate and multivariate analyses aim to describe the static properties of a variable, whereas the aim of a time series analysis is to describe its dynamic properties over time. Knowing an airplane is 10 feet off the ground with the nose angled up and is at full throttle are static variables that would suggest a plane is taking off. However, knowing that over the last five seconds the elevation was 150 feet off the ground, then 140, then 120, then 90, and then finally 60 feet off the ground changes the interpretation of the multivariate data to suggest that the plane is about to crash. The addition of the trend features for the height changes the interpretation of the static data about height, pitch and thrust significantly.

Once there is some level of understanding the data, that understanding can be used to build models that allow for accurate predictions without having to actually measure what is being predicted. The simplest example of this is to use the results of the analysis itself as the model. For example, if someone wanted to predict “heart rate” without having to measure anything, the average heart rate would be a pretty good model. If someone wanted to predict heart rate based on degree of fever, they could use the result of the multivariate analysis that correlates heart rate
with temperature as a model. If someone wanted to predict heart rate based on heart rates during the past hour, they could use the results of a time series analysis as a model. These basic models are frequently employed in clinical medicine, although their complexity is frequently more complicated than is depicted here.

Statistical analysis provides a systematic and standardized process of characterizing data so that it can be understood in the context that it is being analyzed. Modeling endeavors also require a systematic approach, but the range of options is more varied than in statistical analysis(44, 45) since the products of analyses are often used as “building blocks” for a model. It is not uncommon for models to draw on elements from more than one type of analysis in making a prediction. One example of this hybrid technique is the time-course approach to microarray analysis(46, 47). As an example of this approach, the expression levels of twenty different genes are measured to determine their activity in two classes of cancer. If it were to stop here, this would be a basic multivariate model. However, the expression levels of these same twenty genes are measured repeatedly under different conditions and at different points in time. Under the standard multivariate model that used baseline expression levels of the twenty genes, it is impossible to tell which genes determine cancer class. However, by adding the behavior over time in the different nutrient environments, the different classes of cancer can be determined. This is a well established technique for genomic modeling. The technique is based on a paradigm that utilizes time series data elements in a multivariate data format. In multivariate statistical analysis, a high degree of correlation between independent variables (known as multicollinearity – an inherent feature of time series data) can invalidate the results of the analysis by invalidating the calculations relating to the analysis of the independent variables as unique components(41, 42). However, when modeling is focused on the relationship between the dependent variable and
the aggregate of all independent variables (without trying to analyze the independent variables themselves), multicollinearity is permissible(48).

Clinical Prediction Models: For centuries, models have been used to demonstrate our knowledge about the world in which we live. They help us share our understandings about the observations we make, and they help us anticipate what is to come. Models are created, revised, and destroyed based on a constant cycle of probing their parameters and either confirming or disputing their validity based on their ability to make accurate predictions. In astronomy, models allow us to predict where particular constellations will be based on the day of the year and the time of day. Medicine is generally not so precise, partly due to the fact that there is a much larger interdependence between variables, and also because we do not fully understand all of the variables that belong in a model. Nonetheless, thousands of medical models exist and are used to help clinicians interpret and understand the data they are faced with every day. Scoring tools are a type of model that combine multiple data elements, weight them according to their correlation with the outcome of interest, and output a score that can be used in a number of ways. Individual scores can be used to make predictions that can help guide treatment decisions and communications with patients and families. Medical emergency teams use the Early Warning Score(49, 50) or scoring tools similar to it in this fashion. Grouping scores allows standardized comparisons between two or more entities by providing a risk-based adjustment to the outcome of interest(51, 52). For instance, if institution A has a 5% mortality rate and institution B has a 7% mortality rate, it is unclear if the difference is due to differences in institutional performance or if it is due to differences in patient characteristics. Knowing what the predicted mortality rate was in each institution standardizes the measurements by providing a common denominator.
Almost all clinical models are built on multivariate regression or a regression-like approach that evaluates a number of candidate input features (variables) and measures their individual correlation with the outcome of interest. The strength of the correlation is used to assign points for each of the included variables, with more points being assigned for highly correlated variables and for greater deviation from the variable’s normal value. Finally, points attributable to each feature are summed together to provide the composite score that provides an estimate of the net effect of all the features combined.

One example of this type of model is the Pediatric Risk of Mortality Score (commonly referred to as the PRISM score)(10, 11). The PRISM score is used to estimate a child’s mortality risk based on information obtained in the first 24 hours of their admission to a pediatric intensive care unit (PICU). The PRISM score is calculated by providing the most abnormal values for 17 variables during the first 24 hours of care in the pediatric intensive care unit. Depending on the specific values and the child’s age, points are assigned to each variable. To illustrate, a child who has a heart rate of >150 beats per minute (bpm) is assigned 4 points for heart rate. Heart rate is not the strongest predictor of death though – plenty of children admitted to the PICU have heart rates >150 bpm during the first 24 hours and survive. However, if the child’s pupils are fixed and dilated (evidence of severe brain dysfunction), they get 10 points for pupillary reaction: kids that have this degree of brain dysfunction are much more likely to die than those that have a high heart rate – thus the higher score. After points are assigned for each of the variables, all of the points are added together to generate the overall PRISM score. The combined score is then entered into an equation that provides the user with the probability of death during the PICU stay.
Since most of these scoring tools have been built using a multivariate data paradigm that is constrained to a single value per variable, they are generally limited to evaluating a static state at one point in time. They are unable to characterize an extremely important type of information: trends. In order to evaluate a dynamic state over multiple points in time, a time series data paradigm is required. However, since most scoring tools weight their independent variables differently based on regression coefficients, they are prohibited from using data with high degrees of multicollinearity and are therefore unable to use time series data. This is a second limitation above and beyond the fact that doing so would likely violate the requirement for simplicity since these tools are usually used in a manual fashion.

While multivariate models prevail in the setting of clinical prediction tools, there are small but growing number of medical models based on time series data. These models have been used in a number of settings\textsuperscript{(53, 54)} ranging from imputation strategies for missing data\textsuperscript{(55)} to analysis of beat-to-beat variability in heart rate as a way to discriminate survivors from nonsurvivors\textsuperscript{(56, 57)}. However, unlike the multivariate based scoring tools that tend to employ a spectrum of independent variables, most medical models that use time series data have restricted their focus to the time series features of a limited number of independent variable.

Finally, there is the concept of using the results of multiple models as latent independent variables in their own right. While there is precedent for this is in financial and weather forecasting disciplines\textsuperscript{(58, 59)}, it is not a common practice in medicine. There are plenty of examples of studies that compare performance of one model to another, but studies that combine two or more predictive models to arrive at a new prediction are sparse. A general observation noted in our review of these types of studies is that if two or more models are based on similar data, then one of the component models often dominates and there is little effect of adding the
second model. However, if the models are based on disparate data, the resultant model typically performs better than either of the component models in isolation.

*Inpatient Cardiac Arrest as the Example to Illustrate the Process:* In order to build a clinical prediction model that combines the traditional multivariate data elements with the time series data elements, we sought out a problem space that had the following characteristics: 1) target problem has a known relationship to variables measured in a time series fashion; 2) measured variables are abundantly available; 3) time series elements are likely to help predict the target problem. We selected “cardiac arrest in a pediatric intensive care unit” as our target for a number of reasons. First, we were able to identify all cases of cardiac arrest easily since they are recorded on specialized code sheets. Second, standardized criteria(60) can be used to isolate true cardiac arrests from other events that get documented on the code sheets. Finally, cardiac arrest is a significant, life threatening condition that predictably results when a patient’s vital signs deteriorate beyond a point of compensation. As vital signs deteriorate, patients progress from a state of normalcy, to a state of compensated shock, to a state of uncompensated shock, and finally to cardiac arrest. Progressive shock is one of the leading causes of pediatric cardiac arrest.(3) Given that shock can be characterized by vital signs (establishing their plausible association to cardiac arrest) and vital signs are automated and ubiquitously available in pediatric intensive care settings, we felt this was an appropriate example on which to illustrate the process. Furthermore, since shock can often be reversed with treatment, we believe there is a possibility of real world application of the example.

After establishing that cardiac arrest fits the desired criteria, the spectrum of possible conditions that can lead to cardiac arrest must be considered. In reviewing the literature for inpatient cardiac arrest, we determined that patients arrest due to a number of other causes(3),
including intrinsic arrhythmias that can send a patient into immediate cardiac arrest, and unexpected events that can result in cardiac arrest in a matter of minutes, such as sudden uncontrollable bleeding, unplanned removal of life support devices such as ventilators or endotracheal tubes(61), and embolic phenomena such as pulmonary embolism. The list of possible causes is extensive, but almost all causes not attributable to progressive shock share a common feature: they lead to arrest very rapidly. Also from our review of inpatient cardiac arrest literature, we discovered that shock is usually insidious in onset and is characterized by deterioration over minutes to hours, whereas the other causes of arrest are characterized by deterioration over seconds to minutes. Additionally, whereas shock can be well characterized by vital sign data, other causes of arrest are not so well characterized. Given the slower nature of the progressive shock process affords a greater amount of data than the other processes, we felt it appropriate to constrain the example model to parameters that relate to shock.

The fundamental feature of shock is that the body’s need for energy is not being supplied in sufficient quantities. By far, the most frequent cause of shock in the pediatric intensive care setting is one of insufficient oxygen delivery to the tissues(62). Shock can be described from a perspective of supply and demand. On the supply side, oxygen delivery is a process that is dependent on: hemoglobin, oxygen, and blood flow(62). Hemoglobin and oxygen can be measured directly. Measuring a patient’s blood flow, on the other hand, is not commonplace. However, blood flow is a function of heart rate and the stroke volume associated with each heartbeat. Heart rate is measured directly, but again, stroke volume measurements are uncommon. For a fixed vascular resistance, though, stroke volume is proportional to the pulse pressure (the difference between systolic and diastolic blood pressure readings)(63). The pulse pressure can be directly measured. One other nuance regarding oxygen delivery is that it is
dependent on the pressure gradient across the tissue bed, so the gradient between the mean arterial pressure and the central venous pressure is important. Mean arterial pressure can be determined from systolic and diastolic blood pressures. Central venous pressure, on the other hand, is only obtained in a relatively small fraction of the population. Not having this value readily present for the majority of the population is a potential obstacle to being able to model cardiac arrest due to progressive shock.

When examining the variables that relate to the supply of oxygen to the body, most adhere to the desired features of being automatically collected by the monitors, reliably measured, and ubiquitous in the pediatric intensive care population. Oxygen demand depends primarily on temperature and level of activity. Temperature is measured directly, but the method of measurement determines the accuracy of the reading: core temperatures esophageal or rectal probes tend to be more accurate than oral or axillary readings. Furthermore, some measurement modalities are integrated into the physiologic monitoring system, which has two implications: 1) it allows for automated capture, which can also achieved with an electronic medical record; and 2) it allows for continuous measurement, whereas others typically do not. Therefore, care should be exercised when using temperature as a variable to characterize oxygen demand. This introduces another potential obstacle for successfully modeling cardiac arrest. Level of activity is comprised of a host of factors ranging that can include factors such as the work of breathing, digestion, presence of chills and rigors, seizure activity, and a number of other conditions. It is generally not measured in an objective, quantitative fashion, and again, not being able to incorporate it into the modeling process poses a risk to diminishing model performance.
Although we identified at least three potential weaknesses, we have established that there are still a number of variables that are time series in nature and directly relate to the physiology of shock. Despite the risk of not being able to generate an ideal model, we nonetheless felt there was a sufficient amount of data to determine whether the addition of time series data elements incrementally improved model performance as compared to baseline multivariate analysis.

Methods [& Results]

In order use time series data in a clinical predictive modeling paradigm that is based on a multivariate data format we needed to accomplish three fundamental tasks: 1) study models that utilize time series data to perform classification and determine their characteristics; 2) explicitly represent the candidate features that determine the target of interest in both multivariate and time series fashions, including: a) specific measurement modalities; b) windows of observation; c) resolution of observations; and d) computations required to derive the time series features such as slopes and intercepts; and 3) create the modeling data sets using the candidate features in a data structure supported by the modeling algorithm.

The method we are proposing is listed below as a series of steps. In order to maintain continuity of focus between the method and the results, we will begin each section by identifying the task and providing a general description of the concepts and theories that we are applying. As the result for each step, we illustrate the step using the specific case of modeling cardiac arrest in a pediatric intensive care unit. [The illustration is indented and placed in brackets.]

Determine Model Characteristics: Time series data are used in models from a variety of disciplines. In order to draw from the existing techniques, they must be examined and their properties characterized. Starting at the broadest level, we initially searched web-accessible
articles for “time series” and “prediction model.” A basic exploration of the qualitative properties of the resultant hits produced several observations that provided focus for subsequent analyses.

[The first observation was that some models rely on raw statistical associations while other models utilize explicit equations for mathematical or physical properties. For instance, financial models tended to have a more statistical focus while engineering models tended to provide mathematical representations for the phenomena being studied. The second observation was that the majority of models utilize information from past events. This measure of the seasonality features of time series data is germane to many areas of medical predictive modeling, but it does not apply to cases where initial or singular events are the target, which is the case in this study. The final observation was that “pattern recognition” and “classification” tasks more precisely describe the focus of our study.]

Refining our screening query to “time series” and either “pattern recognition” or “classification” we obtained a more homogenous group of studies, including a greater fraction from medically related fields.

[However, clinical models were still lacking, and strategies to predict initial events were rare. One class of studies that seemed to hold promise was based on time course analysis. Frequently used in genomic classification tasks, this strategy has been used to classify (in the initial trial) diseases where less than 100 samples were available for training but thousands of candidate variables were being analyzed. The method relies on defined variables (gene expression levels) under defined conditions (exposure to different agents) at defined times of measurement (baseline and several post-exposure times).]
methods share many of the properties we desire for modeling cardiac arrest, and we therefore focused our efforts on adapting this strategy as a template for our work.]

Select Candidate Features: Selecting the set of variables to serve as candidate features that will discriminate between different classes is of key importance in modeling. A combination of approaches, including literature review, mind mapping, and statistical analysis are several methods that can be used to identify plausible features. Ideally, features should describe the target, correlate with the target, or have other plausible associations with the target.

[Since shock states are characterized by imbalances between supply and demand, and since neither is constant, there is no established variable or combination of variables that can be used to identify the threshold at which to define shock. However, we can define a set of variables that semiquantitatively represent supply and demand, and we can measure several markers of anaerobic metabolism, which takes place during shock states where demand exceeds supply. In addition to the direct determinants of shock, there are associated variables that may modulate the baseline risk of cardiac arrest: the overall metabolic profile, comprised of various salts in the body is one modulator, and the functional status of the organs of the body is another. These modulators are often measured as laboratory values. In order to select candidate features for modeling cardiac arrest, we generated a list of variables based on mind mapping and literature review that are relevant to shock and matched them with data that was electronically available. The mind map can be found in Figure 1 and the resulting list of variables can be found in Figure 2. Note that several of the variables identified in the mind map were not electronically available.]
**Select Measurement Tools:** Often times, candidate features can be measured in a variety of ways. Recall the example given above for temperature: it can be measured by digital devices, old fashioned mercury thermometers, or temperature sensitive chemical strips and it can be obtained from the skin, various mucous membranes, the tympanic membrane, the temporal artery, or even from the bloodstream. It can also be recorded continuously or intermittently. Defining how candidate features are to be measured and are not to be measured helps in understanding potential strengths and weaknesses of the models being built.

[Several candidate features for cardiac arrest prediction were measured by multiple means. Heart rate can be measured by EKG signals or by pulse oximetry. Blood pressures can be measured continuously by arterial lines or intermittently by blood pressure cuff. Temperature can be measured by all the methods described above. As a general rule of thumb, when presented with multiple possibilities, we selected the measurement modality that had the highest reliability. For heart rate, we selected EKG signals. Even by allowing for multiple means of measurement, temperature availability as an electronic source was present in <10% of the population, so we had to exclude it from the list of candidate features. Blood pressure determinations posed a particular problem, though: there is a difference in data resolution between continuous arterial line readings and intermittent noninvasive (blood pressure cuff) checks. Clinically, when these two measurements disagree, neither is uniformly more accurate than the other. Also, arterial line pressures are not ubiquitous in the population. From a pure “desired properties” standpoint, we felt the noninvasive measurements were more ideal since they are obtained on everyone and since they don’t require a procedure to obtain. However, since blood pressure is so fundamental to the concept of shock, and since arterial line tracings provide a more]
detailed representation of blood pressure, we felt it appropriate to include both modalities in the model.]

*Standardize Candidate Feature Formatting:* Time series data are different from multivariate data in one fundamental way: they are repeated measures rather than a single measurement. In order to make use of either data type, its properties need to be understood and a strategy for transforming its native properties into the desired properties of the candidate features needs to be developed. The net result of these steps is to provide explicit formatting specifications for each of the candidate features. The steps to this process are shown graphically in Figure 3.

*Determine Class of Representation:* The first step is to determine whether the feature should be represented in a multivariate format (single value) or in a time series format (multiple values). To make this decision, one must evaluate the tradeoff between potentially useful trend information in the time series format and the complexity it adds to the modeling process. Time series data can be collected in fixed intervals (such as vital signs in physiologic monitors) or they can be collected in nonstandard intervals (such as laboratory measurements). Fixed intervals are somewhat easier to manage, since the predictable timing between measurements allows for consistent representation between subjects. Nonstandard intervals pose the problem of many measurements being taken in a given time period for some subjects with single or no measurements being taken for other subjects. Nonstandard intervals can still be represented in a time series format, but additional specifications need to be determined for how to standardize their representation. Another strategy is to encode nonstandard interval features in a multivariate format, using single values (first, last, mean during some timeframe…) to represent them in the modeling process.
[We represented physiologic monitor data in a time series format and laboratory and demographic data in a multivariate format. Both noninvasive blood pressure measurements and laboratory measurements are characterized by nonstandard intervals of measurement. We treated the noninvasive measurements as time series data since their frequency of measurement far exceeded the frequency of laboratory measurements. Many patients in the arrest group had multiple laboratory measurements taken in the hours preceding their arrest, and we were concerned that representing these as time series elements could bias model performance, since the number of unique measurements taken could itself serve as a feature distinguishing arrest from control. Although this could be viewed as a legitimate feature, we felt that the risks imposed by the operator controlled nature of this variable and the potential bias that it would introduce in isolating “time series effects” outweighed the benefits of using it as a feature unto itself.]

**Identify Reference Point:** The second step is to identify a reference point for the time series features so that their relationship to the target of interest is standardized. Typically this will be a particular event (such as cardiac arrest) and measurements can be referenced by the number of minutes that they preceded the event. This strategy would typically be employed in situations where the candidate features lead to the event, which is also the target. If the reference point is an event that may lead to changes in the features, and the target is something else, then measurements can be referenced by the number of minutes that they follow the event.

[We selected the cardiac arrest event as the reference point in time and represented candidate features by the number of minutes that they preceded the arrest.]

**Specify Windowing Parameters:** The third step is to constrain the time series features to a specific window of time and to specify the resolution of measurement within the window. At this
step, higher resolution (more frequent) measurements are generally preferable to low resolution measurements, since subsequent steps will use multiple data points in calculations, and higher resolution provides for a more accurate representation of the underlying data than a lower resolution. However, provided there are enough measurements specified in the chosen window to accurately represent any trends of interest preceding (or following) the reference point, the resolution can be reduced from the native resolution of the measurement tool. The resolution does not need to be uniform for the entire window. If trends of interest occur close to the reference point, then measurements taken closer to the reference point can be represented in a higher resolution and those that are farther from the reference point can be represented in a lower resolution.

[Based on our understanding of the physiologic changes that precede an arrest, we chose to include measurements that were taken up to 12 hours prior to the arrest. However, changes in vital signs in the hour before the arrest tend to be more rapid and pronounced than changes that occur greater than an hour before the arrest. In particular, the most dramatic changes occur in the 10-15 minute window before an arrest. For this reason, we chose a resolution of every one minute for vital signs taken in the one hour window preceding the arrest, and we chose a resolution of every one hour for vital signs taken in the 12 hour window preceding the arrest.]

**Transforming Native Properties to Desired Properties:** The final step required to standardize the formatting of candidate features is to transform the native measurement resolution to the resolution specified by the windowing parameters. When the number of native measurements in a given time period preceding or following the reference point exceeds the desired number of measurements specified, a reduction strategy is needed. Several options include selecting the
mean, median or mode, maximum, or minimum of the native measurements. When the number of native measurements is less than the desired number of measurements specified, an imputation strategy is needed. Several options include imputing a normal value, a mean, median, or mode value from the data, or carrying forward previous data points.

[We selected the latest measurement taken prior the arrest to represent our multivariate features. The time series features were already represented by every one minute measurements, so no additional transformations were necessary for the one hour window preceding the arrest. Noninvasive blood pressure measurements were an exception. Measurements of noninvasive blood pressures ranged from every one minute to every 60 minutes. Since the corresponding counterparts (arterial line measurements) were continuously measured, we imputed missing values by a simple carry forward strategy and imputed a predefined normal value when no prior measurements were available to carry forward. We recognize that more sophisticated imputation methods are available, but considered this level of detail beyond the scope of the current project. For the 12 hour window preceding the arrest, the 60 native measurements (every minute) were averaged to obtain the single value specified by the windowing parameter.]

Compute Latent Variables: Although simple inclusion of the raw time series features may be sufficient to proceed to building a model for cardiac arrest, there are explicit computations that may further improve the ability of the models to discriminate one class from another. At least two such types of computations exist: trend features in the time series data, and clinically relevant latent variables that represent concepts not directly encoded in the raw data. The steps to this process are shown graphically in Figure 4.
Clinically Relevant Variables: When two or more of the features can be mathematically combined to express another feature that has an association with the chosen target, explicitly performing the calculation and encoding the new feature as a latent variable may help discriminate between two target classes. Although many of the advanced modeling algorithms may inherently be able to properly classify without the explicit calculation of the clinically relevant latent variable, it may serve as a way to minimize the size of the modeling data set by allowing for elimination of the core features used in their calculation.

[For the cardiac arrest case, two candidate features, the shock index and the oxygen delivery index, can be calculated from the variables we selected for our analysis. Given that the calculated measures often convey a better representation of shock than any single variable in isolation, we thought it prudent to explicitly include these two latent variables in the modeling process. Shock index is determined by dividing the heart rate by the systolic blood pressure. Oxygen delivery is approximated by multiplying heart rate, pulse pressure (the difference between systolic and diastolic blood pressures), oxygen saturation, and hemoglobin. Since each of these values is dependent on continuously measured parameters, we managed them according to the continuous measurement data paradigm. Of note, at least one other clinically relevant feature would have theoretical utility in a cardiac arrest model: oxygen extraction index - the difference between arterial and venous oxygen saturations. If blood supply to a tissue bed diminishes, oxygen delivery diminishes, and the amount of oxygen extracted from the arterial blood increases, thereby decreasing the amount of oxygen in the venous blood. Although there is theoretical utility to this measure, obtaining the arterial-venous oxygen difference requires a central line and unless the central line has a venous oximeter, the measurement
requires two simultaneous blood gas measurements. We were unable to include this variable in the model since the data to perform this calculation was rarely present in the raw data set.]

**Trend Features:** For each feature, its time series data is represented graphically as the value of the measurement plotted against the resolution of time specified in the windowing parameter. Depending on the nature of the data and the specific trends that one would expect to help distinguish one class from another, candidate trend features should be explicitly encoded as numerical representations by performing the computations necessary to characterize the features of interest. This can include any number of representations, but the slopes and intercepts, and the mean values for various intervals of time relative to the reference point are standard methods of characterizing “trends” and overall status for any given interval. Unless the trends of interest are uniform and precise, a strategy of calculating slopes, intercepts, and means for multiple intervals may provide a better chance to discriminate between classes than a single set of calculations. Additionally, beyond the single determination of slope, intercept, or mean for any given interval, expressing combinations of features, such as the ratio of the mean of one interval compared to the mean from another interval, may provide an even better discrimination.

[Since the trends leading to a cardiac arrest are not well characterized and vary from case to case, we derived multiple permutations of slopes, intercepts, and means for several prearrest intervals. We chose a 5 minute prearrest interval since the majority of arrests demonstrate a more pronounced deterioration in vital signs during this interval. We also chose 10, 15, and 60 minute prearrest interval in order to provide a diverse representation of trends occurring during multiple prearrest intervals. We also explicitly represented ratios between mean measurements for each interval to each of the intervals]
that preceded it: 5/10, 5/15, 5/60, 10/15, 10/60, and 15/60. Finally, considering that steeply negative slopes combined with baseline low means may provide the best measure of discrimination, we divided the slope by the mean for each prearrest interval in an effort to maximize the contrast of such a combination.]

**Seasonality Features:** Seasonality features of time series data also hold important information that is not characterized by multivariate data or by trend features of time series data. If such features are of interest, then mathematical representations of their properties are required if they are to be used as representative features. These may come in the form of autoregression and moving average models (ARIMA) or any number of other time series analyses. As in the case of trend features, if the seasonality features are not uniform and precise, multiple analyses may be necessary in order to properly characterize the seasonality features and encode them as latent variables.

[In the case of predicting cardiac arrest, we are only interested in identifying the initial event, not on predicting future arrests based on previous arrests. We therefore did include any features related to seasonality for the purpose of this study.]

*Create Modeling Data Subsets:* The final step in preparing to generate predictive models from time series data is to determine what actually belongs in the model. The effort that was put into selecting candidate features, specifying their properties and format, and deriving additional clinically relevant and trend features has provided a comprehensive set of candidate features that are suitable for modeling, but all may not be necessary or appropriate for modeling. From this comprehensive set of features, any number of data subsets can be instantiated to answer the question of “Did that step help improve the model?” In order to answer the question, two or more subsets of the comprehensive data set need to be evaluated through a formal modeling process
that involves training and tuning on one data set, and measuring their performance in another. Relative performance from one reference set to the other serves as a measure of the utility of the candidate features that differ between them. The candidate modeling data subsets are shown in Figure 5.

[Our first task was to construct a data set to serve as a baseline for how well multivariate data alone discriminated arrest cases from control cases. For this baseline data set, we included measurements and physiologically relevant computations that were closest in time to the arrest event. None of the time series features were included in this data set. The second task was to add the raw time series data elements to the baseline data set. The purpose of this model was to determine whether or not the addition of time series elements, without their associated trend calculations, changed the model accuracy. The third task was to add the clinically relevant latent variables with the time series elements to determine whether or not the latent variables provided any additional value to classifying arrest versus control. The fourth task was to add the time series trend features to the baseline data set (without the raw time series elements) to determine whether or not the trend features changed the model performance. The final task was to construct a data set containing all available elements. The purpose of this model was to determine if the combination of time series data elements and their abstracted trend features changed model accuracy beyond any changes that may have occurred by either in isolation.

Discussion

Multivariate models serve as the basis for almost all clinical prediction tools that currently exist. Most of these tools operate on a single value per variable paradigm and are incapable of
factoring change over time into their results. Even if the tools called for specific calculations of change over time measures, they would most likely not be used since most of these tools require manual data entry. We have presented a methodology that allows models to incorporate time series data so that change over time can be factored into their results. In doing so, we have identified a number of practical implications.

We believe automation is necessary in order for any time series based model to be successful and sustainable. Multivariate models and tools can be cumbersome to use, even with only a handful of data points to enter. The number of data points in time series approaches prohibits manual use. Identifying which candidate features to include in the modeling process needs to balance the usefulness of the feature in discriminating between model targets with the ease of getting the feature’s measurements into the model. Data feeding the model needs to be populated automatically if the model is to be simple enough to be accepted. Manual data entry of large amounts of data is impractical for sustained model use. These factors need to be considered as specific measurement tools are sought for each of the desired features being included in the mode. Also, the candidate features should be fairly ubiquitous in the study population and part of routine patient care. Although this isn’t an absolute requirement, having readily available data facilitates model development by providing a rich retrospective body of data on which to build and tune the model. Data that is rarely obtained has a greater chance of biasing model performance, and will serve to limit the population to which the model is applicable. Likewise, data that requires specialized equipment to obtain or is not part of routine patient care increases the complexity of the overall process since it would likely prohibit a prospective protocol for data collection – again potentially biasing model performance since consenting an entire population of patients is unfeasible. Once data can feed into the model automatically, filtering
and processing the data into a standardized format is necessary. The initial determination of the format needs to be informed by a conceptual model that relates the candidate features to the targets of interest. Once the format is determined though, extensive preprocessing must occur in order to transform the raw measurements for each variable into its desired format. The methods chosen to accomplish these preprocessing steps should also be compatible with automation.

The addition of time series analysis results as latent candidate features opens up a realm of possibilities – only a few of which we’ve covered in this study. We’ve specifically presented possibilities of using calculations based on clinically meaningful concepts and trend analysis. However, if targets of interest occur repeatedly, then analyses of seasonality features can also be included as candidate features. To date, these types of analyses have received attention in isolation, but have not been used as candidate features with a broader set of variables in a comprehensive theoretically based modeling construct as we have proposed here. This is work to be done.

While current models and tools are typically based on regression strategies with multivariate data, the quantity of data involved in modeling with time series elements requires algorithms that are more robust when ratios of cases to candidate features is small. These tools have made some progress into clinical literature, and several clinical tools are based on them, but their use has still been limited to a multivariate data paradigm. Historically, tools have relied on black box models have encountered significant resistance for clinical use, since the mechanisms used to determine their results were hidden. As more studies show these tools to be superior to their regression based counterparts, we anticipate these advanced algorithms will gain broader acceptance, making clinical acceptance of the methods proposed here more likely.
Finally, we have demonstrated application of this methodology in a case of cardiac arrest prediction in a pediatric intensive care setting. We hope we have presented each of the steps with sufficient detail that they can be generalized to other scenarios where historical data is used to derive information that helps guide actions. Being able to automatically analyze this type of data and integrate it with other features to arrive at a final determination has inherent potential as decision support that can help end users such as bedside caregivers understand the information contained in the historical data.

**Conclusions**

Reviewing predictive modeling paradigms from nonclinical domains provided a potential template for incorporating time series data elements into clinical prediction tasks. The techniques used for time course analysis, which uses patterns of measured gene expressions over time to distinguish different biological classes, can be adapted to analyze repeated clinical measures in a similar fashion. We have proposed a six step method for creating clinical prediction models using time series data. The steps include: developing a theoretical construct around a prediction task; selecting candidate features and specifying their properties to match the theoretical construct; preparing data for modeling by transforming it from its raw state into a standardized representation; characterizing clinically relevant calculations and time series analysis as latent candidate features; and creating data subsets to measure the effect that each category of candidate feature has on the predictive accuracy of the resulting model.

We have used progressive shock as a common mechanism leading to cardiac arrest in a pediatric intensive care unit as an illustrative example. A successful prediction model for this phenomenon based on automated techniques could be used to monitor patients continuously for
the risk of cardiac arrest. Continuous, objective measures are more likely to systematically identify these children than are the intermittent checks that characterize how they are currently identified. Implementation of such a tool could ultimately serve to save the lives of hundreds, potentially thousands, of children every year in the United States alone.

**Acknowledgements and Funding**

This work was partially supported by a NIH K22 Early Career Development Award: 5 K22 LM008389. James Turley, RN, PhD was the Chair of the academic mentorship committee that provided project oversight. Jack Smith, MD, PhD, Noriaki Aoki, MD, PhD, Hualou Liang, PhD, and Michele Mariscalco, MD were also members of the academic mentorship committee. All members of the academic mentorship committee were funded through their standard academic appointments. Betty Ligon, PhD provided draft proofreading and revision service and was also funded by her standard academic appointment.
Figure 1: Determinants of shock include variables that represent supply, demand, measures of end organ function, and modulators of the metabolic environment. Clinically useful concepts that are not explicitly measured but directly relate to supply and demand of oxygen are represented as explicitly calculated latent variables.
Figure 2: Physiologic variables selected as determinants of shock. Each variable has been matched to a source that contains the raw measurement. For measurements with more than one source, the preferred source was used for the match.
Figure 3: Multivariate variables were always assigned as the most recent measurement taken before the reference point: the event in this illustration. For continuously measured variables, a multivariate representative was assigned according to the multivariate parameters. The remaining elements were considered time series elements: 59 minute by minute measurements in the one hour preceding the event, and hour by hour measurements in the 12 hours preceding the event.
Figure 4: Latent variables were derived from the raw physiologic Multivariate and Time Series data sets. Clinical Latent Variables were based on calculations used in clinical medicine in assessing for shock. Trend Analysis Latent Variables were based on slopes, intercepts, means, and the ratios of these features for 5, 10, 15, and 60 minute windows that preceded the arrest event.
Figure 5: Five candidate modeling subsets of data were created to determine the impact of time series and trend analysis latent features (separately) to baseline multivariate model accuracy. Clinical latent variables were compared to multivariate + time series features to determine their relative impact to model accuracy. Finally, all candidate features were combined to determine the net impact of time series + clinical latent + time series latent features on model accuracy.
References


44. Berry MJA, Linoff G: *Data mining techniques: For marketing, sales, and customer relationship management. (Second Edition)*. Indianapolis, IN, Wiley 2004


46. Ebert BL, Golub TR: **Genomic approaches to hematologic malignancies**. *Blood* 2004;104:923-932


49. Edwards ED, Powell CVE, Mason BW, Oliver A: **Prospective cohort study to test the predictability of the cardiff and vale paediatric early warning system**. *Arch Dis Child* 2009;94:602-606


61. Mort TC: **Unplanned tracheal extubation outside the operating room: A quality improvement audit of hemodynamic and tracheal airway complications associated with emergency tracheal reintubation.** *Anesth Analg* 1998;86:1171-1176


64. Jensen BN, Jensen FS, Madsen SN, Løssl K: **Accuracy of digital tympanic, oral, axillary, and rectal thermometers compared with standard rectal mercury thermometers.** *Eur J Surg* 2000;166:848-851
Building Clinical Prediction Models Using Time Series Data
Modeling Cardiac Arrest in a Pediatric ICU

Curtis Edward Kennedy, M.D., M.S.¹
James P. Turley, Ph.D., R.N.²

¹Baylor College of Medicine, Department of Pediatrics, Section of Critical Care Medicine
One Baylor Plaza, Houston, TX 77030
²The University of Texas Health Science Center at Houston – School of Biomedical Informatics,
7000 Fannin Suite 600, Houston, TX 77030

Corresponding author
Curtis Kennedy, MD, MS
cekenned@texaschildrenshospital.org
Baylor College of Medicine
Department of Pediatrics
Critical Care Section
One Baylor Plaza
Houston, TX 77030
832-826-6230
Building Clinical Prediction Models Using Time Series Data: 
Modeling Cardiac Arrest in a Pediatric ICU

Abstract

Background
Every year thousands of children die after experiencing a cardiac arrest while being cared for in a pediatric intensive care unit. In acute care units, scoring tools help identify sick children who are likely to arrest. In pediatric intensive care units (PICUs), where children are far more likely to have an arrest, no such tool exists. These tools that do exist do not account for variables changing over time and, therefore, are unable to recognize deterioration. We recently proposed a theoretical construct for incorporating time series analysis into a predictive modeling task. The focus of the current study is to describe explicitly the modeling process and to address factors that arise from using time series data.

Methods
Starting from a point where candidate features and a target have been determined, we present sequential steps and considerations relating to time series data required to train and validate clinical prediction models using time series data. Given that time series models are too complex for manual implementation, we identify the steps that require automation in order to implement the models being generated in a real world use scenario.

Results
Of the seventeen modeling steps we present, nine require automated techniques in order to achieve real world use. Although the initial modeling effort requires a significant amount of manual effort, once the key processes required for subsequent use are automated the models can
operate continuously without any further user input – a key improvement over most current models.

Conclusions

Creating and validating predictive models using time series data rely on the same fundamental processes that are employed for modeling in standard multivariate data. We have presented a process that can be automated in order to accommodate prospective use of these models in clinical environments and have demonstrated the steps of the process in an example of predicting cardiac arrest in a PICU. Although we have illustrated the process involved in predicting a pediatric cardiac arrest, it needs to undergo trials in other clinical settings to determine if it is robust as a general solution to using time series data in clinical prediction modeling.

Introduction

The practice of medicine involves the interpretation of a “wide and deep” diversity of data to synthesize meaningful information that can be used to provide care to patients. As the practice of medicine is becoming increasingly data-driven, the need for tools that help provide a standardized approach to the interpretation and use of various types of data has grown. Thus far in clinical medicine, the tools that have been developed have been able to address the “wide” aspect of data by being able to aggregate data from multiple sources into prediction models or scoring tools that function to determine some net effect or event probability. These tools almost always rely on a single representation for any given variable being used and, therefore, have been unable to leverage useful information that is contained in the “deep” aspects of the data – namely its change over the course of time.
We believe that incorporating changes in data over time into clinical tools holds great promise for increasing their reliability and validity, and we have recently proposed a methodology to encode the time series (TS) aspects of frequently measured variables into a data format that can be used in a number of modeling algorithms to create clinical prediction models. Although subsequent processes for actual model generation and validation are well-established, they are not all utilized in all modeling scenarios. Additionally, because the complexity of models based on time series (TS) data is greatly increased compared to the standard multivariate models that are already in existence, we consider it appropriate to describe explicitly the methodology from start to finish.

Many of the steps involved in modeling are a field of study unto themselves. They also have more than one alternative approach to achieving the desired outcome. As we describe the steps necessary for model training and validation, we will present a brief overview of the purpose of the step, some competing strategies if applicable, and our justification for selection. Some of the steps are specific to model development, whereas others also apply to subsequent use of a model. For steps that are required for subsequent use, we considered the ability to automate the step an absolute necessity. On occasion, this resulted in a degree of theoretical detriment to the performance of the associated step, but we feel that this tradeoff is necessary if these types of models are ever to achieve clinical use.

**Methods [and Results]**

In order to maintain continuity of focus between the method and the results, we will begin each section by identifying the task and providing a general description of the concepts and
theories that we are applying. As the result for each step, we illustrate the step using the specific case of modeling cardiac arrest in a PICU. [The illustration is indented and placed in brackets.]

The training and validation of predictive models is a multistep process that includes merging data from different sources, preprocessing the data to remove noise, imputing missing elements, and normalizing the data for use in modeling algorithms, partitioning the data for specific uses, and, ultimately, generating and validating the models. An assumption to begin these steps is that the candidate features being considered in the model have already been determined. For the case of predicting cardiac arrest in the PICU, we have listed the variables we determined by methods presented in our previous work(1).

Identify Cases and Controls: In order to build a prediction model, the target on which it will train must be valid and reliable, and example cases must accurately represent the range of phenomena being predicted. When trying to discern differences between cases and controls attributable to a single factor, controls are paired with cases by matching (2) on the remaining factors that are potentially confounding. Although this matching process minimizes the differences attributable to the confounding factors, in prediction modeling it serves to introduce bias into the model (3, 4) because it restricts the representation of normal variability and other non-matched features in the training process. Ultimately, a balance is needed between the benefits of matching cases and controls and the benefits of explicitly providing unmatched, purely random controls.

[We operationally defined cases as patients who received more than two minutes of external cardiac massage for acute physiologic deterioration. We determined this definition in an effort to screen out a large number of acute deteriorations that were secondary to events, such as endotracheal tube occlusions and unplanned extubations.]
The mechanism leading to arrest in these situations is fundamentally different from our chosen target: arrests caused by progressive shock. Additionally, monitor alarms typically notify bedside caregivers of the acute deteriorations, whereas the arrests caused by progressive shock typically are more insidious in onset. We selected controls from two populations. The first was one that was completely random and represented the baseline status of pediatric intensive care populations: random bed space with a randomly determined reference time. The second population included subjects who were matched with cases based on age, sex, diagnostic category, and highest severity of illness as determined by the PRISM-III score (5). The reference time for this population was determined by the worst set of vital signs obtained in the 24-hour period that the PRISM-III score represented. The second population represented the patients with progressive shock who did not progress to cardiac arrest. We felt that including both populations maximized the possibility that the models would be able to discriminate impending cardiac arrest from random physiologic variability and from otherwise equally ill but stable patients. This step is germane only to model training and validation, so there is no need for automation.

Of note, the proportion of cases to controls at the time the model is generated needs to be close to 1:1 for most of the modeling algorithms (6). Populations that are disproportionately represented in the training and calibration data set put the resulting models at risk of being biased (7, 8). There are two main ways to achieve this final relationship of cases to controls: under-sampling the majority class (typically the control population), or over-sampling the minority class (typically the case population) (9, 10).
[Because we had a reasonably large population of cases and because there seems to be some evidence that under-sampling may be preferable to over-sampling(11), we chose to under-sample the population of control cases. This step also is germane only to model training and validation, so automation is not needed.]

Identify Data Sources: In order to populate data into fields specified as candidate features, the sources for the data must be identified. Frequently, a given data element can be obtained from multiple sources. In selecting sources to provide data, two questions should be considered: 1) Is one source more accurate than the others? and 2) Can the total number of sources be reduced by choosing to obtain data from a particular source? Most models initially are built using existing data sources, primarily because obtaining large amounts of data quickly is easy and the expense and time required to collect new data for modeling are prohibitive. Also, prospective data collection in human subjects, when the data is intended for research purposes, requires obtaining informed consent of all potential subjects, which is unfeasible in a setting in which the data needed for the study must be collected before the patients have been identified as meeting criteria to become subjects.

Using existing data collected for operational purposes allows data from an entire population to be used because protected health information can be removed from the data before they enter the research realm, providing that elements of the protected health information are not candidate features.

[In order to obtain the data for the candidate features we specified in our model, we required data extracts from three sources: a clinical data repository that contained demographic and laboratory data, a physiologic monitor archive that contained vital sign
and pulse oximetry data, and handwritten ‘code sheets’ that contained details for all cardiac arrest cases that occurred in the PICU during the study period. This step is germane only to model training and validation, so automation is not needed.]

*Merge Data:* Once the data sources are identified, a method to match records from one source to another is required. The exact method for doing this will vary from one case to the next, depending on the formats that the source data are in. Typically, a common key such as a medical record number, or a common set of keys such as name and birth date, will provide enough detail for matching records. The key or keys need not be the same for all sources, but each must be sufficiently unique so as to provide a single, exact match in each source.

[In our study, this task was a challenge because each data source had at least one key used for matching fields that was dependent on a manual process. The code sheets had handwritten date/time and bed space fields. The data for admission, discharge, and transfer (ADT) transactions in the data repository contained manually entered timestamps. The physiologic monitors required manual entry of patients’ names into the monitor. The manual processes occasionally resulted in a discrepancy of several hours among sources. In order to overcome this challenge, the data extraction team first matched the clinical data repository with the code sheets based on medical record numbers. They then matched the code sheets to the physiologic monitor archive based on bed number and date/time. They then had to perform a manual synchronization procedure in order to align the three data sources. Accurately matching records required the team to manually review the data and triangulate information such as admissions, discharges, bed space changes, and name changes in the physiologic monitor archive. This process would not be a viable strategy moving forward due to time and resource
requirements. We have identified the changes that are necessary in order to link directly the clinical data repository and the physiologic monitor archive, and we would need to implement these changes before we could use these data sources to feed a model in a prospective fashion. This step is first one requiring automation because matching the records between the data sources is required in order to properly populate the features the models will use.]

Identify a Reference Time: In order to standardize the TS relationships of the candidate features, a common reference point that is shared among all subjects needs to be defined. Numerous options are available to accomplish this task and include admission, discharge, or transfer transactions, or a specific event such as a cardiac arrest. Unless all of the data sources share a common source for date and time, synchronization of date and time among the different sources (that contain time-sensitive data) is a necessary precursor to this task (see above).

[We selected the cardiac arrest event as our reference time and defined the TS candidate features according to their relationship with the arrest event. In our case, our data sources were not synchronized, resulting in a probabilistic distribution of recorded dates and times around a reference time. Because cardiac arrest was our outcome of interest and the code sheet specifically recorded the date and time associated with the event, we utilized this source as our reference. After merging records from the code sheet with records from other data sources, we adjusted the times in other data sources to the time recorded on the code sheet. The data fields from the physiologic monitor archive were TS in nature, and the arrest event was operationally defined in the physiologic monitor log during a 30-minute window around the designated date and time of the arrest as one minute before the first time the heart rate dropped below 50 beats per minute
(bpm), or to its minimum if it clearly dropped but did not reach 50 bpm. If criteria for heart rate were not met, pulse oximeter readings of < 70% (or to its minimum if it clearly dropped but did not reach 70%) were used as the secondary criteria. If neither heart rate nor pulse oximeter criteria were met, then a systolic blood pressure drop below 50 mmHg (or to its minimum if it clearly dropped but did not reach 50 mmHg) was used as the tertiary criterium. This determination was usually but not always obvious, so consensus among the three members of the data extraction team was obtained for each case. We selected the heart rate as the principal variable because it is the standard criteria by which an arrest is defined. We chose pulse oximeter as our secondary variable because it was measured on a minute-by-minute basis. Drops in blood pressure were our tertiary variable in cases for which acute changes in heart rate and pulse oximetry were not identified. Dates and times from the clinical data repository were fairly easy to separate into “before” and “after” categories relative to the cardiac arrest reference date and time provided from the code sheet, as all elements that were time-sensitive originated from a laboratory system that maintained a date and time (different from the manually entered ADT transaction date and time data) that were already synchronized with the physiologic monitor archive. We defined the representative value for the time-sensitive laboratory parameters as the most recent value obtained before the arrest event. Although we employed highly manual processes to identify the reference time and to synchronize times in our data sources, this step is germane only to model training and validation. Prospective use would utilize “now” as the reference point and enter TS elements relative to it – regardless of whether or not the specific date and time stamps agree among various
data sources. Therefore, in this case, we do not consider routinely automating reference point identification and synchronization among data sources as necessary.]

Window Data: Once data sets have been merged and dates and times synchronized, the data must be constrained to the period of time determined during the planning stage when specifications for each candidate feature were established. This step is straightforward: remove data that were obtained before or after the established period of time.

[For the cardiac arrest data set, we windowed the data set to include data for 12 hours preceding the arrest event. Windowing the data is required for subsequent model use, so we recommend automating this process. However, only the removal of data before the established period of time is necessary to automate.]

Migrate Data: In many circumstances, the tools used to extract and merge data sets are not the same tools that will be used to train and validate models. Even if each of the tools stores its data in a common database, the specific data sets used for modeling often differ from those used for extraction and merging. This implies that translation between the two (or more) data sets needs to take place. Because several preprocessing steps are required to produce the final data set for model generation and validation and because the preprocessing steps alter the raw data, we consider the safest (not the most efficient) approach is to create multiple intermediate data sets, each of which contains the output of one of the preprocessing steps.

[In the cardiac arrest case, we migrated from a MS SQL Server 2000 platform to a MATLAB (Release 7.0a) platform, copying the original data elements in a text delimited format. We selected this strategy because of concerns of network bandwidth requirements that would be associated with linking MATLAB to the MS SQL Server – copying allowed us to perform subsequent manipulations using local resources. Whether this step
is required depends on the relationship between the modeling tool and the database. Because this process is required for subsequent model use, if migration is required then it should be automated.]

_Detect and Remove Outliers:_ One of the first tasks of data preprocessing is to identify and remove data that are likely to be erroneous. No method is perfect, and all methods will fail to detect truly erroneous data and will misclassify some valid data as erroneous (12). Outlier detection can be done by pure statistical methods (13), by establishing allowable thresholds based on prior knowledge (14) of an acceptable range, or by several more sophisticated approaches that utilize modeling techniques similar to those discussed in this study (15). Of note, outlier removal changes the nature of the data by removing naturally occurring outliers in addition to the erroneous readings that are the intended target. Some experts advocate retaining the outliers and leaving the modeling algorithms responsible for offsetting their impact (16, 17).

[Pure statistical methods rely on numerous assumptions about the pattern of data distribution and the frequency of outlier observation (18). Because these assumptions are unlikely to hold across a broad spectrum of physiologic data that ranges from normal distributions to highly skewed distributions, we utilized limit-based detection using a conservative set of parameters for each of the candidate features (Table 1). Because no evidence exists for optimum limits, we assigned relatively arbitrary limits using the following heuristic: “the highest value, or the lowest value, above or below which values outside the range would be clinically equivalent to the selected value.” Although more sophisticated methods of outlier detection are available, they are relatively complicated compared to limit-based detection and are unlikely to automate as easily. We acknowledge the perspective of not removing outliers. However, we considered it
plausible to exclude data that fell outside the ranges we specified because they were clinically equivalent to their corresponding boundary and were more likely to be measurement errors than statistical outliers. Outlier detection and removal are required for subsequent model use, so automation of the step is necessary.

**Impute Data:** Because the algorithms that are used to create models require that each case have an identical data format to all other cases (19), missing data fields are not permissible. Missing data can be imputed by numerous methods (20, 21), including allowing null values to populate the field, carrying forward previous values, substituting either mean or median values for the variable, or using more sophisticated imputation strategies based on trends in the surrounding data points.

[We chose two different strategies, depending on which type of variable was being imputed. For data that had historical values, we used carry forward imputation, copying the most recent value forward into the empty field. We chose this strategy for two reasons: 1) the assumption “if it wasn’t measured then it hasn’t changed” is a fairly conservative, commonly employed approach; and 2) automating this imputation strategy is far simpler than are the alternative approaches. In situations in which the data being imputed did not have a previous value to carry forward, we imputed a predefined normal value for the variable. We selected an arbitrary “normal” rather than “mean” or “median” because these measures of central tendency generally were not normal (reflecting the average global illness of critically ill patients), and we considered it more appropriate to use an assumption of “if it wasn’t measured then it was normal” rather than creating new abnormal data based on a statistical distribution from an inherently skewed population.}
Imputation of data is required for subsequent model use, so automation of the step is necessary.

Reduce Data: Given the requirement for an identical data format for each case, extra data fields also are not permissible. Extra data can be reduced by a number of methods, including using the highest, lowest, first, last, mean, median, mode, or a given percentile of the multiple values. Mean and median allow for data to be reduced to a more central measure, whereas highest and lowest values allow for data to be reduced to more directed boundary conditions. The relationship of the candidate feature being reduced to the target of interest and the number of values being consolidated in the reduction step should help determine the appropriate method of data reduction. If boundary conditions are desired, using a designated percentile close to the boundary edge may help minimize the chance of unfiltered outliers being selected. Of course, once the step has been performed, the raw data that were used to determine the new value should be removed from the data set.

We reduced our data by averaging the values of the minute-by-minute vital signs for one-hour blocks starting at one hour before the reference point and ending at 12 hours before the reference point. We considered that averages more accurately represented the physiologic status during the one-hour blocks than did any of the other measures, and the averaging task is simple enough to implement in an automated fashion even though it is not the simplest of the available methods. Once the averages were obtained, the raw data used to calculate the averages were removed from the data set. Data reduction is required for subsequent model use, so automation of the step is necessary.

Calculate Latent Features: Latent variables are used to encode new candidate features (that are not directly measured) from raw features present in the data set (23-25). When there are
concepts that may better represent a relationship to the target than does any single measured feature in isolation, and if the concept can be derived through mathematical operations on the raw features, then latent features can be calculated and added to the modeling data set.

Determining whether or not to include latent features, and how to calculate them, should be undertaken during the planning stage when specifications for each candidate feature are being established.

[Several clinically useful concepts can be expressed mathematically and are more sensitive to determining "shock" than is any single vital sign in isolation. We selected two of these calculations to include as candidate latent features in the modeling process: shock index (SI) (26, 27) and oxygen delivery (28) index. SI is defined as heart rate (HR) divided by systolic blood pressure (SBP): SI = (HR / SBP), and can be directly calculated from existing features. The oxygen delivery index (ODI) that we calculated was an approximation based on values that typically are derived through use of pulmonary artery catheters, which are used infrequently in pediatric intensive care. Oxygen delivery is defined as the product of cardiac output (CO) and oxygen content of blood (CaO2): ODI = (CO * CaO2). Calculating the precise oxygen content requires the arterial oxygen pressure, hemoglobin (Hgb), and oxygen saturation (SPO2) values. It frequently is approximated by multiplying Hgb and SPO2 because under most circumstances the fraction contributed by the dissolved gas (arterial oxygen pressure) is negligible. We used this approximation to represent oxygen content of blood in our oxygen delivery index: CaO2 = (Hgb * SPO2). Because we did not have cardiac output as a directly measured feature, we had to use its definition: the product of HR and stroke volume (SV): CO = (HR * SV). We were able to measure HR directly, but we had to approximate SV through]
use of the Frank-Starling principle that correlates SV to pulse pressure (PP) (28). Using this principle to make this approximation assumes a fixed vascular resistance (which is reasonable for short windows of time but does not necessarily hold over long periods of time). Combining these relationships, our operational definition for the ODI was: ODI = (HR * PP * Hgb * SPO2).

Of note, we included the use of ODI as a latent variable based on its theoretical utility as a trend feature to discriminate between arrest cases and controls. Its change relative to previous values was the feature of interest. We are not advocating for its authenticity as a reliable measure of oxygen delivery. Additionally, we would have included other theoretically useful latent features, but we were unable to do so because the raw data necessary to perform the calculations were not commonly available. One such feature was the oxygen extraction index – the difference between arterial and venous oxygen saturation. As supply decreases or demand increases, the amount of oxygen extracted from the arterial blood supply increases, resulting in a lower oxygen saturation in the venous blood. Calculating this feature requires simultaneous measurements of arterial and mixed venous blood gas samples, or simultaneous oximetry measurements from arterial and venous sources. Raw measurements for each of these features were very limited in the study population, so we were unable to include the oxygen extraction index as a candidate feature. Finally, if latent features are required for subsequent model use, automation of the step is necessary.

Calculate Time Series Latent Features: Explicit analysis of the TS data can be used to create additional latent features. Two classes of information can be derived from the TS data: trend features and seasonality features (29, 30). Trends can be represented mathematically as
regression equations that output slopes and intercepts. Seasonality features can be represented mathematically by any number of TS analyses, including frequency-domain and time-domain methods. We consider this step to be a key factor in using TS data for predictive modeling. We are presenting a very limited and focused example in this study, choosing to address only the trend aspects of the TS data. The range of potential analyses that could be used in creating TS latent variables is vast. As before, determining which TS latent features to include, and how to calculate them, should be undertaken during the planning stage when specifications for each candidate feature are being established.

[Even though we limited our scope to trend analysis, the optimum trend features to calculate have not been established. Furthermore, although the greatest rates of change in the pre-arrest state occur in the 5-15 minute window preceding the arrest, the timing of onset and the rate of deterioration vary substantially among patients. In order to maximize the likelihood of finding an optimum set of trend parameters, we selected four pre-arrest intervals and calculated means, slopes, and intercepts for each interval: 5, 10, 15, and 60 minutes. In addition to the straightforward calculations, we explicitly calculated ratios of means between each interval and each larger interval: 5/10, 5/15, 5/60, 10/15, 10/60, 15/60. Finally, because the combination of mean and slope was potentially more discriminating than was any given trend feature in isolation, we maximized the contrast in this relationship by dividing the slope by the mean for each pre-arrest interval. Again, because subsequent models will need to use them as determining features, calculating the TS latent features should be an automated process.]

Normalize Data: Once the data set has been properly formatted, cleansed by outlier removal and imputation, and latent variables have been calculated, the data should be normalized so that
the modeling algorithms can perform optimally (22). Many algorithms will automatically normalize the data, if it has not already been performed, but intentional normalization provides more control over the process and will prevent problems due to improper or absent normalization during model training. The goal of normalization is to bring all variables to a common scale, such as 0 to 1, so that the modeling algorithms are not overly influenced by values that are measured on larger scales. Numerous strategies are available to normalize data (22) and range from common min-max and z-score strategies to more mathematically complicated strategies such as regression-based techniques (31). Furthermore, for any given strategy, additional normalization parameters need to be specified. Examples include: 1) how the normalization operation should be parsed: every variable versus families of variables versus the entire data set; and 2) whether or not each normalization operation should result in a Gaussian distribution (32) (in addition to resulting in a range from 0 to 1).

We selected a common min-max strategy because it has been shown to be effective for many types of data and it is easy to automate. Min-max normalization is a straightforward calculation of: \( x' = (x - \text{min})/(\text{max} – \text{min}) \) (19). We normalized TS features in family aggregates: HRs were normalized as a whole, as were SBPs, diastolic blood pressures, pulse oximeter readings, and so on. Multivariate variables such as pH and lactate were individually normalized. Although most modeling algorithms perform at their best when working with normally distributed data, we were not convinced that performing further mathematical operations on the normalized data set in order to approximate a Gaussian distribution would provide more benefit than risk, so we did not perform this step. However, many variables already exhibited a Gaussian distribution, so this decision likely was trivial. The normalization procedure is the final step that needs to
be automated for subsequent model use. The steps that follow are specific to model training and validation.]

Create Candidate Modeling Data Sets: Because of the number of new classes of data that have been added to the traditional multivariate paradigm, a method to distinguish the effect of each class on model performance is needed. The number of new classes can vary, depending on the granularity of detail that is desired. Examples mentioned so far in this study include raw TS features, concept-related latent features, and TS-related latent features. Each of these broad classes could be divided further to determine effects attributable to individual families of variables, such as HR, blood pressure, and pulse oximeter. To measure the effect(s) on model performance attributable to a class of features, a cycle of model training and validation in a reference set of features should be compared to a cycle of model training and validation in a set of features that include the class of features being considered.

[We created five classes of candidate modeling data sets: 1) multivariate; 2) multivariate + raw TS features; 3) multivariate + raw TS features + clinically relevant latent variables; 4) multivariate + trend-based latent variables; and 5) all combinations of data. Training and testing models from these five sets allow for the measurement of relative contributions of raw TS features, the clinically relevant latent variables such as SI and ODI, the trend-based latent variables such as slopes and intercepts of various time intervals preceding the arrest, and, finally, the net effect of combining all elements. Comparing accuracy between data sets that underwent feature reduction to those that did not allows for the measurement of over-fitting effects attributable to the high number of candidate features.]
Partition Data Sets: Once the data have been normalized, they are ready for modeling. Although the entire data set could be used to train a model, the ability of that model to perform in unseen data (its external validity) would be undetermined. Models should not be put into use until their external validity has been determined (33). There are two ways to establish external validity. The gold standard is to employ the model in a prospective fashion and measure its prediction accuracy over time, noting any degradation that may occur as a result of changing practices or disease prevalence. However, this is a laborious process and typically is not undertaken unless some measure of how well it is likely to perform already exists. Therefore, in order to estimate the external validity, a random portion of samples usually is withheld so that they are never accessed during the training process (22, 34). Determining how many samples to withhold involves a tradeoff: withholding too many samples may degrade the model’s performance, whereas withholding too few samples may inaccurately estimate the model’s external validity. Depending on how many examples are in the data, between 20 and 50 percent of the samples typically are withheld for validation as the final step in assessing model performance.(35, 36)

[We randomly withheld 33% of our data as a validation set. Because we had only 103 cases on which to train, and more than a thousand candidate features, we decided that the training set needed to have as many cases as possible. We did not go as low as 20% because we were concerned that numbers of cases would be insufficient to provide a sufficient resolution of accuracy measurement during model testing: one case in 20 equates to a 5% difference, whereas one case in 33 equates to about 3%.

Reduce Candidate Features: Many modeling algorithms are able to discriminate even the subtlest differences between case and control classes of data. When the number of examples is
small and the number of variables that the model can use is large, modeling algorithms are prone
to memorize the classes (8). If this occurs, the preliminary results of the model accuracy will be
superb, but the model will fail to predict classes accurately in the validation data set. In order to
prevent this problem, feature reduction frequently is employed prior to training to reduce the
number of candidate variables to a number less than the algorithms need to achieve
memorization. As a field of study, feature reduction is broad and has literally hundreds of
described techniques that span numerous underlying strategies (37). At a very basic level,
though, feature reduction can be performed by starting with no features and sequentially adding
them, or it can be performed by starting with all features and sequentially removing them.
Additionally, it can be performed in concert with a modeling algorithm, or independently.

[We selected two strategies that had proven performance in microarray analysis:
recursive feature elimination (RFE), which removes features independent of the modeling
algorithm, and support vector machine weighting (SVMW), which adds features based on
weights assigned during SVM model training. Given that both of these techniques are
sensitive to interdependencies among variables for a given outcome (i.e., they will retain
two or more variables that work together to make a prediction, even if neither in isolation
is highly correlated with the outcome), and that the modeling tool we were using
(MATLAB Spider) supported these techniques, we chose them over the simpler,
correlation-based techniques. In order to estimate the degree of over-fitting, if any, we
also trained candidate modeling data sets without employing feature reduction and
compared their performance to those that were trained with feature reduction.]

Train Models: Once all of the previously described steps have been performed, various
modeling algorithms can be used to train models. Before doing so, though, one other point
regarding data set management is necessary. So far we have identified two data sets: one for training and another for validation (external validity). Nomenclature in the literature often identifies these as training, testing, and validation data sets (38). In our methods, we are treating the training and testing data sets a single data set because the modeling tool we have chosen to use can automatically parse and reparse this data set multiple times as necessary to create its own internal representations of training and testing data sets. Many modern modeling tools have this ability to manage training and testing sets automatically from a single data set through use of “cross-validation” routines. These routines essentially split the data set into internal training and testing sets, generate a model and test it internally, and then repeat the process a number of times (known as N-fold cross validation) (38, 39). In repeating this process multiple times using slightly different data sets for training and testing, the models converge on an optimum set of parameters that work the best across multiple splits. This process also generates a measure of internal validity that characterizes how consistent the models are between splits. If models have chaotic internal performance during the training phase, it is somewhat risky to rely on their performance in unseen data as a truly reliable measure of external validity.

As was the case for feature selection, an abundant number of strategies and algorithms can be employed to generate predictive models (40). Again, at a very basic level, models can be divided into two groups: those that operate with explicit variables and rules and those that operate in a black-box fashion, in which the specific mechanism for determining the model output is hidden. To date, there is no single best way to create a predictive model, so several strategies commonly are employed in a trial-and-error fashion in order to identify the optimum strategy for the data at hand. Historically, tools used in clinical medicine have relied on regression-based methods, which work via explicit variables and rules. Decision trees are a class of predictive models that
share the same features of explicitness but are based on the more sophisticated algorithms that characterize modern modeling strategies. Although models that use explicit variables and rules provide a degree of understanding to the user as to how a particular prediction is determined, they often do not achieve the same level of raw performance as models that operate in a black-box fashion. Two types of modeling strategies that have proven to be robust classifiers in multiple domains are neural networks and support vector machines (41).

[Not knowing what the optimum modeling strategy was, we used a combination of candidate modeling algorithms for each of the candidate data sets. For explicit rule based classifiers, we built linear regression and decision tree models. For black-box based classifiers, we built neural network and support vector machine models. Each of the modeling strategies had numerous parameters that could be specified as additional tuning parameters. For the purpose of this study, we accepted default parameters because our goal was to compare various combinations of features and core modeling strategies. Further improvements in performance may have been obtained with an in-depth exploration of parameter effects, but this was considered to be beyond the scope of this study.]

*Validate Models:* For any given model that is produced, its performance must be measured in order to demonstrate its utility. Although we alluded above to the distinction between internal and external validity measurements, we did not discuss the specific measures used to assess model validity. As was the case with many of the previous steps, this is an active area of research, and multiple methods of model assessment have been described (8, 42).

[For each model generated, we measured classification accuracy, sensitivity, specificity, and the area under the receiver operating characteristic curve (AUROC).]
These measures of accuracy are well-established and accepted across multiple disciplines, and we considered them to be sufficient to demonstrate differences among the various models we generated. Of note, the ratio of cases to controls in our data set did not reflect the true prevalence of cardiac arrest (the ratio had to be balanced in order for the modeling algorithms to perform optimally). Therefore, positive predictive values and negative predictive values, although listed in performance reports, were meaningless. As a result, other measures that rely on these values (such as gain and the F-statistic) were also not reported in our results.

Discussion

Building clinical prediction models based on TS data elements is technically equivalent to building models based on standard multivariate data elements. However, whereas most resulting multivariate models can be translated into tools that can be used manually, those built on TS data elements are too complicated to both translate and present data to, even if they could be easily translated. As a result, all of the steps necessary to acquire raw data, clean it, and otherwise prepare it for modeling need to be automated. This is a significant departure from the vast majority of clinical scoring tools. Numerous examples of tools that are based on TS data exist, but we have not found any that approaches a given problem from a comprehensive theoretical perspective and leverages information from each potential feature that can help discriminate between two classes of targets. Multivariate-based scoring tools that are manually implemented can use data from flowcharts, monitors, physical exam, laboratory systems, and other means to produce a score. Current TS-based scoring tools typically use a single channel, and even if they are multichannel, they are still constrained to a single source system. Our proposed method
blends these two strategies to create a physiology-based model for cardiac arrest, and it assimilates TS data from multiple data sources in order to build and run the model.

The method we have detailed in this study illustrates from start to finish the entire 17-step process, starting with identification of cases and controls and ending with validation of the final models. We have assumed that the theoretical construct for the model, including the list of candidate features and specifications for their representation, has already been determined. If it has not, we have described that process in a previous study (1). Of the 17 steps that we describe, nine are required for both model development and subsequent use and, therefore, require an automated strategy for their implementation. The remaining eight steps are specific to model development and do not have this requirement.

The method we have presented follows methods that are commonly practiced in predictive modeling. Specific options available for any given step in the method create an array of possible modeling endeavors that is too vast to cover in a single study. We have attempted to provide examples of commonly employed options and the factors that should be considered when selecting an option, and we have presented our choice along with its justification. It is unlikely that all of our choices were the most optimal solution, and we recognize that there are many ways to achieve a goal. We expect that active, engaged discussion and debate about rationales for choices ultimately will cast better light on issues that are sure to exist but have not been elucidated in this study.

Ultimately, our hope is that these steps will encourage other similar clinical predictive modeling endeavors and that the class of prediction tools that use TS data in making their determinations continues to evolve. Historical data is relevant and useful in making predictions.
Current multivariate models simply cannot leverage it. These studies contribute to the ongoing determinations to meet the need.

**Conclusions**

We have presented a series of 17 steps that are required to build a clinical prediction model based on TS data. Unlike typical traditional multivariate tools that can be used manually, TS-based tools will require automated processes for handling data in order to be used prospectively. We have identified eight key steps that require automation and have provided recommendations on ways to achieve it. This is the first example of a clinical prediction model being built based on TS data elements that were determined by referencing a broad theoretical construct for the model target. We have illustrated the process in a case of modeling progressive shock as a mechanism leading to cardiac arrest in a PICU.

**Acknowledgements and Funding**

This work was partially supported by a NIH K22 Early Career Development Award: 5 K22 LM008389. James Turley, RN, PhD was the Chair of the academic mentorship committee that provided project oversight. Jack Smith, MD, PhD, Noriaki Aoki, MD, PhD, Hualou Liang, PhD, and Michele Mariscalco, MD were also members of the academic mentorship committee. All members of the academic mentorship committee were funded through their standard academic appointments. Betty Ligon, PhD provided draft proofreading and revision service and was also funded by her standard academic appointment.
Table 1: High and low limits to identify outliers for removal from physiologic and laboratory data sets. When no values were available to carry forward for imputation, “normal values” from the Imputation column were used.
References


10. Liu AY: The effect of oversampling and undersampling on classifying imbalanced text datasets. A thesis for a Master of Science in engineering. Austin, TX, The University of Texas 2004


32. Hua J, Balagurunathan Y, Chen Y, Lowey J, Bittner ML, Xiong Z, Suh E, Dougherty ER: 
   **Normalization benefits microarray-based classification**. *EURASIP J Bioinform Syst Biol* 
   2006:43056

33. Tropsha A, Gramatica P, VK: **The importance of being earnest: Validation is the absolute** 
   essential for successful application and interpretation of QSPR models. *QSAR & Combinatorial 

34. Ian H. Witten, Frank E: *Data mining: Practical machine learning tools and techniques*. 
   Amsterdam, Netherlands, Morgan Kaufman Publishers 2005

35. Stein R: **Benchmarking default prediction models: Pitfalls and remedies in model** 
   validation. *Journal of Risk Model Validation* 2007;1:77-113

36. Altman DG Royston P: **What do we mean by validating a prognostic model?** *Stat Med* 
   2000;19:453-473

37. Guyon I, Elisseeff A: **An introduction to variable and feature selection.** *The Journal of 
   Machine Learning Research* 2003;3:1157-1182

38. Hsu C, Chang C, Lin C: **A practical guide to support vector classification.** Available 

39. Kohavi R: **A study of cross-validation and bootstrap for accuracy estimation and model** 

40. Lim T, Loh W, Shih Y: **A comparison of prediction accuracy, complexity, and training** 
   time of thirty-three old and new classification algorithms. *Machine Learning* 
   2000;40:203-228

41. Byvatov E, Schneider G: **Support vector machine applications in bioinformatics.** *Appl 
   Bioinformatics* 2003;2:67-77
Using Time-Series Analysis to Predict Cardiac Arrest in a Pediatric Intensive Care Unit

Curtis Edward Kennedy, M.D., M.S.¹

James P. Turley, Ph.D., R.N.²

Noriaki Aoki, M.D., Ph.D.²

M. Michele Mariscalco, M.D¹

¹Baylor College of Medicine, Department of Pediatrics, Section of Critical Care Medicine
One Baylor Plaza, Houston, TX 77030
²The University of Texas Health Science Center at Houston – School of Biomedical Informatics,
7000 Fannin Suite 600, Houston, TX 77030

Corresponding author
Curtis Kennedy, MD, MS
cokenned@texaschildrenshospital.org
Baylor College of Medicine
Department of Pediatrics
Critical Care Section
One Baylor Plaza
Houston, TX 77030
832-826-6230
Using Time-Series Analysis to Predict Cardiac Arrest in a Pediatric Intensive Care Unit

Abstract

Objectives: To build and test cardiac arrest prediction models in a pediatric intensive care unit using a modeling framework that incorporates time series features, and to measure changes in prediction accuracy that were attributable to different classes of time series data elements.

Methods: A retrospective cohort study of pediatric intensive care patients over a 30 month study period. All subjects identified by code documentation sheets with matches in hospital physiologic and laboratory data repositories and who underwent chest compressions for two minutes were included as arrest cases. Controls were randomly selected from patients that did not experience arrest and who survived to discharge. The modeling data set was based on twelve hours of data preceding the arrest (or reference time for controls).

Measurements and Main Results: 103 cases of cardiac arrest and 109 control cases were used to prepare a baseline data set that consisted of 1025 candidate features comprised of four data classes: multivariate data (MV), raw time series data (TS), clinically relevant (CR) latent variables, and the derived trend characteristics (TRD) for the time series data. We trained a matrix of models for five combinations of data subsets and four modeling algorithms, and measured the performance of each model in a validation data set not used for model training. The MV model using the regression algorithm had a baseline accuracy of 78% and 87% area under the receiver operating characteristic curve (AUROC). The MV+TRD model using the support vector machine algorithm had the best performance, with accuracy of 94% and 98% AUROC.
The highest ranking features in the peak performing model were evenly split between MV and TRD data classes.

**Conclusions:** Cardiac arrest predictions based solely on a multivariate data and a regression algorithm misclassified cases 3.7 times more frequently than predictions that used derived trend features and a support vector machine algorithm. Although the final model lacks the specificity necessary for clinical application, we have demonstrated how information from time series data can be used to increase the accuracy of clinical prediction models.
**Introduction**

Children are admitted to pediatric intensive care units (PICU) to receive the highest level of attention while they are being treated for life-threatening diseases. The population is inherently unstable, and patients change rapidly between states of improvement and deterioration. When deteriorations happen, it is incumbent on the bedside caregiver to detect the deterioration, assess its potential impact to the patient, and intervene if necessary. Each of these functions is operator-dependent, so patients receive different levels of service at different points in time. Despite continuous monitoring of their vital signs and high staffing ratios, thousands of children suffer cardiac arrest events in PICUs every year (1-4). Many of these arrests are preceded by deteriorations in the patients’ vital signs (5-8), suggesting that progressive shock may contribute to the arrests.

In an ideal environment, cardiac arrests attributable to undertreated shock could be avoided by early recognition of deteriorations that precede an arrest and timely intervention before it occurs. The goal of this study is to focus on the recognition aspect of the problem by evaluating the utility of using time-series (TS) data to characterize deterioration (a time-dependent phenomenon) as input into a prediction model for cardiac arrest in a PICU. Models that use TS data to characterize trends preceding an event differ from the vast majority of traditional clinical prediction models. The first difference is that the number of variables required to use the model is orders of magnitude larger for TS models than for traditional models. The second difference is that TS models require more complicated data preprocessing to be performed before they can be used. These differences have implications that could impact the adoption of models using TS data in actual clinical use. Therefore, a secondary goal of this study is to identify potential barriers associated with using TS models and to suggest strategies to overcome them.
Cardiac arrests are associated with mortality rates typically in excess of 60% and high
disability rates in survivors, so significant effort has been made for their prevention and
management (9-15). Antecedents to cardiac arrest identified in the literature have been described
primarily in the context of patients in acute care units who deteriorate to the point of requiring
transfer to an ICU or an arrest (16-18). These studies suggest that deteriorations often are
detectable hours before arrests occur and that patients often are evaluated beforehand but fail to
receive treatment adequate to prevent the event. These findings have led to the implementation
of medical emergency teams (METs) in most institutions. In conjunction with METs, scoring
tools have been developed and deployed to help objectively assess patients for risk of having
life-threatening deteriorations (19, 20). These tools have been widely implemented, but their
target population is one of relatively healthy patients, and their purpose is to differentiate a
patient who is sick from one who is healthy. ICUs contain a population of patients who have
been determined to be sick, so the scoring tools that have proven useful in an acute care setting
are unable to identify the patients who are most likely to suffer cardiac arrest. Furthermore, even
if the tools could be recalibrated, the border between compensated and uncompensated shock
differs from patient to patient and even at different times for the same patient depending on
factors such as activity and temperature. Because the risk threshold is a moving target, clinicians
often use data from earlier points in time to interpret new data and determine its implications for
the patient. Tools that can perform these interpretations automatically are needed.

Due to the variable nature of the threshold between compensated and uncompensated shock,
the physiologic monitors to which patients are connected are not good tools for identifying
patients at risk of having cardiac arrest. In fact, these monitors are plagued by such a high rates
of false-positive alarms that nurses frequently ignore the alarms (21). Patients and families even
learn how to silence them. Several types of tools have been developed that help minimize false-positive alarms and are intended to provide clinicians with more accurate predictions of deterioration. One type relies on a multichannel strategy that calculates deviation from normal across multiple physiologic measurements and amplifies its risk prediction as functions of both degree of deviation and number of channels shown to be deviating. Another type relies on statistical process control charts that automatically recalibrate their limits and alert for deteriorations that result in limits falling below a threshold value. Both of these technologies have accuracies superior to those of standard physiologic monitors because they require more than a single episode of abnormality. Monitors are designed to be extremely sensitive in detecting single indicators of life-threatening conditions, but their sensitivity carries a price of poor specificity. However, the former method lacks the ability to utilize data from trends over the course of time, and the latter method lacks the ability to integrate multiple simultaneous channels. The value of merging these two modalities for purposes of physiologic monitor alarming has not been reported.

We recently proposed a theoretical framework for generating clinical prediction models using TS data. Theoretical advantages of this framework reach beyond simple merging of statistical process control and multichannel analysis. The foundation of the proposed framework is comprised of two elements. The first is standard to most modeling endeavors: formally characterize the mechanisms leading to the event of interest and identify specific variables to include as candidate features in the model. The second explicitly calculates trend features in the TS data (such as slope and intercept) that precede the event of interest and includes them as candidate features in the model. The current study is designed to test the proposed framework by applying it to the case of cardiac arrest prediction in a PICU.
In order to develop a prediction model, the appropriate data must be analyzed by statistical algorithms that are capable of detecting differences in the data between two or more targets of interest. Most scoring tools in clinical use are based on models that were built using multivariate (MV) data structures (19, 27-31), for which single values represent a variable of interest, and relatively simple regression-based algorithms. Although these algorithms perform well in smaller data sets, they have not proven robust in large data sets in which the number of candidate features far outweighs the number of training examples. Microarray analysis presents such a challenge, and more sophisticated statistical algorithms have been developed that overcome the limitations of regression-based analysis (32, 33). Neural networks (NNs) and support vector machines (SVMs) are two examples of modeling tools that use these more sophisticated algorithms and have been proven to be robust in larger data sets (34). One of the limitations of these two modeling tools, though, is that they operate in a black-box fashion: the rules for how a particular output is reached are intangible. One of the virtues of regression-based analysis is that the output is determined by an explicit, tangible mathematical equation. Decision trees (DTs) are a class of models that have proven robust in data sets that contain large numbers of candidate features (similar to NNs and SVMs), but they are able to explicitly output their rules (similar to regression).

In order to accomplish our primary goal of evaluating the utility of TS data in building a prediction model for pediatric cardiac arrest, we aim to determine: which of the candidate input features serve as the best predictors of cardiac arrest and which modeling algorithm generates the best predictive model. The methods we present below to accomplish these tasks have dependencies that must be satisfied before models can be either generated or subsequently used. These dependencies are presented as discussion following our results.
Methods

Setting: This is a retrospective cohort study of patients admitted to a tertiary care PICU in a metropolitan area serving a referral population of over 4 million people. The study protocol was submitted to our institutional review board and received its approval. We identified 103 cases of cardiac arrest that occurred in the PICU by reviewing code sheets that were generated when patients received acute, intensive resuscitation between July 2006 and December 2008. These dates were determined by availability of physiologic monitor logs. Criteria for inclusion as an arrest case were: 1. event location in PICU; 2. first cardiac arrest in the PICU; 3. external cardiac massage for at least two minutes; and 4. able to be matched with records from the hospital’s data repository and from the hospital’s physiologic monitor database. We also identified 109 control cases from patients admitted to the PICU by random selection from the following three categories: 1. first six hours of admission (representing the most consistent point in time when patients experience rapid change and receive multiple interventions); 2. day of maximum severity of illness when score occurred on or after day two of hospitalization (representing deteriorations after admission); and 3. random point in time during hospitalization (representing baseline noise in PICU physiology). Criteria for inclusion as a control case included: 1. did not experience a cardiac arrest in the PICU; 2. survived to discharge (to exclude deteriorations that occur before death in patients who have active orders not to attempt resuscitation [DNAR]); 3. was able to be matched in the physiologic monitor database and the data repository; and 4. was selected by random number generation to be included as a control case (limiting the number of control cases and keeping the ratio between case:control roughly equal to satisfy assumptions of the modeling algorithms). The following methods are high-level abstractions of the steps we took to accomplish the study. In depth details for each step have been reported in prior works(35). In
general, the specific strategy chosen for any given task was selected based on several criteria that included feasibility for automation and historical familiarity (i.e., we chose to use simple, proven strategies that perform well over complicated or esoteric strategies that may have performed somewhat better).

Data: Specifications for an initial data set (Figure 1) required merging data from the three data sources: code sheets, data repository, and physiologic monitor database. In addition to matching records from the three data sources, arrest cases required additional time synchronization among the data sources. Times reported on the code sheets were entered into the monitor database for each case, and arrests were identified using heart rate, then pulse oximeter, and finally blood pressure criteria if neither of the first two variables identified the arrest event. The data repository and monitor database times were already synchronized. The minute before the arrest event was defined as the reference time for arrest cases, and control case reference points were assigned randomly within the designated block of time that defined their respective category (first six hours of admission, the day of their maximum severity of illness, or purely random).

A filtered data set was extracted from the initial data set by selecting only data elements that preceded the reference point by 12 hours (TS data elements) and the final measurement preceding the arrest (MV data elements). Because we chose MATLAB as our modeling platform, and because it was not part of the hospital’s core infrastructure, we generated a de-identified data set from the filtered data set by removing HIPAA defined patient identifiers. The resulting modeling data set was exported from the hospital’s database system to a fixed width text file that we used for subsequent analysis in MATLAB.
**Preprocessing:** We performed outlier removal and imputation in the modeling data set using a limit-based, carry-forward strategy. When no values were present from which to carry forward, a normal value predetermined for each field was imputed. These steps resulted in each case having a value for each variable. For each of the TS elements, we averaged the minute-by-minute values by one-hour blocks for each of the 12 hours preceding the reference point and discarded the minute-by-minute TS elements for all values that preceded the reference point by over 60 minutes. This resulted in a data set in which TS elements were represented by 60 high-resolution data elements (every minute for one hour prior to the reference point) and 12 low-resolution data elements (hourly 12 hours prior to the reference point). We then performed latent variable calculations to determine explicitly the shock index (heart rate / systolic blood pressure) and an oxygen delivery index (heart rate * pulse pressure * hemoglobin * % oxygen saturation) for each corresponding set of vital signs. The shock index is a value reported in the literature that can be calculated directly from raw measures\(^\text{36}\). The oxygen delivery index is based on the oxygen delivery equation used in hemodynamic calculations\(^\text{37}\), but because cardiac output is not measured directly, we used heart rate and pulse pressure as surrogate variables.

Before proceeding to the preprocessing steps that prepare the data for modeling, we performed statistical analysis and visualization of the modeling data set, comparing arrest cases to control cases. The goal of this step typically is to identify variables that are likely to discriminate cases from controls and justify their inclusion into a regression analysis that assigns a relative weight to their relationship with the target being modeled. However, we did not use this step to exclude variables as candidate features because our modeling framework was based on predetermined physiologic principles and the effects attributable to the TS analysis were
undetermined. Instead, the purpose of this exercise was to help visualize and understand the data we were providing to the models as candidate features.

Next, we performed min-max normalization on each category in the data set (i.e., we normalized all TS elements in a given variable as a single block). The final preprocessing step was to partition the data set into a training set (67% of the data), and a holdout validation set (33%) was used to test model performance in data that were unseen during the training phase.

**Modeling:** Because the modeling framework we were testing had not been tested previously, both the optimum set of variables on which to base the model and the optimum modeling algorithm were undetermined. The main study objective was to determine if TS elements could improve baseline model accuracy, but there are two main categories of elements: raw variables and derived trend features. Additionally, it was unclear if explicit representation of clinically relevant calculations would impact model performance. In order to test each of these factors, we constructed five independent modeling data sets: 1. pure MV (utilizing the single measurement closest in time to the point of reference); 2. MV+TS (utilizing all raw measurements); 3. MV+TS+CR (utilizing all measurements, adding explicit representations of clinically relevant calculations); 4. MV+TRD (utilizing the single measurement closest in time to the point of reference, adding the derived trend features from the TS data, but not using the raw TS data elements); and 5. MV+TS+CR+TRD (utilizing using all raw measurements, adding explicit representations of clinically relevant calculations and derived trend features from the TS data). (Figure 2)

In order to determine the optimum modeling algorithm, each of the five training data sets was presented to each of the four candidate algorithms for model training: 1. linear regression (LR); 2. j48 decision tree (DT); 3. neural network (NN); and 4. support vector machine (SVM). Each
of the candidate algorithms could be further tuned beyond its default performance by varying various modeling parameters, but the resulting number of possible permutations that would result from such an exercise was considered to be beyond the scope of this study. For each combination of data set + modeling algorithm, we performed 10-fold cross validation in order to estimate the standard error associated with each algorithm. We then generated one representative model for each combination using the entire training data set. The models then were presented unseen data from the validation data set, and the resulting predictions (arrest versus control) were analyzed to determine: 1. overall accuracy (ACC); 2. sensitivity (SN); 3. specificity (SP); and 4. area under the receiver operating characteristic curve (AUROC). Because the data sets contained a balanced ratio of case:control in order to satisfy modeling algorithm assumptions, other accuracy measures that relied on positive predictive value (or precision) in their determination (e.g., the F1 score and gain measures) were not performed.

Memorization and over-fitting are potential problems associated with data sets that have relatively small numbers of training examples and large numbers of candidate features. Although measurement of performance in an unseen data set helps ensure that a model’s performance is not an effect of over-fitting, feature reduction frequently is employed to reduce the number of candidate features prior to the model generation step. In order to assess for effects of over-fitting in the data set that include all potential candidate variables and in the otherwise best performing data set, we performed two classes of feature selection: SVM weighting (SVMW) and recursive feature elimination (RFE). In order to determine the optimum number of features, we varied the number of allowed features from 15 to 50. We repeated the training and testing steps detailed above and compared the new accuracy measures to the former measures in order to estimate any over-fitting effects.
The final step was to perform a qualitative assessment of the results we obtained in the steps above and to determine if any general patterns were evident in the results. For this step, we analyzed mean performance characteristics for each type of modeling data set across each of the modeling algorithms and for each of the modeling algorithms across each of the modeling data sets. We also probed into the model weights to determine which variables were considered by the different modeling algorithms to be the most predictive features, looking in particular for features that were conserved between algorithms.

Results

Our specifications for defining arrest cases identified 103 cases of initial cardiac arrest events that occurred in the PICU and had corresponding data in the physiologic monitor database and the data repository. From the subjects who met inclusion criteria, 109 controls were randomly selected. Seven code sheets could not be matched, due to illegibility (5) and to inability to match in the corresponding data sets (2).

Lab variables had 0.47% of their data identified as outliers, and physiologic variables had 1.18% of their data identified as outliers. Imputation accounted for 16.8% of lab data and for 1.7% of physiologic variables.

Statistical analysis of the modeling data set demonstrated differences in mean values with a p-value of < 0.05 for: 1) 16 of 20 (80%) MV features; 2) 413 of 497 (83%) TS features; 3) 155 of 288 (54%) CR features; and 4) 182 of 220 (83%) TRD features. Figure 3 shows the differences in heart rate, oxygenation, and systolic blood pressure for arrest versus control cases. For arrest cases, with the notable exception of respiratory rate, each of the vital sign categories demonstrated drops in mean values starting as many as 20 minutes from the arrest, with more
drastic drops occurring in the 5-minute window before the arrest. The shock index and oxygen delivery index encoded as CR latent variables also demonstrated acute worsening before the arrest event, with the oxygen delivery index being somewhat more evident due to starting its decrease approximately 10 minutes before the shock index.

Internal performance measures in the training data set using 10-fold cross validation yielded accuracy measures ranging from 51% to 90%. MV data yielded the poorest performance, with a mean internal accuracy of 66% ± 4%. The best mean performance was measured in the MV+TRD data, with a mean internal accuracy of 79% ± 4%. The TRD features increased mean internal accuracy by 13% (p<0.0001) compared to the MV model. Raw TS elements added 3% ± 3% to the accuracy of the MV model (p = NS). The CR elements (shock index and oxygen delivery index) increased accuracy 1 ± 3% (p = NS). Combining all variables together, accuracy was reduced from the peak accuracy reference (MV+TRD) by 6% ± 5% to 73% (p=0.02), demonstrating a modest over-fitting effect. Compared to the MV models, only models that included TRD features demonstrated a statistically significant increase in mean accuracy.

With respect to internal performance measures of the modeling algorithms, LR yielded the poorest mean performance with a mean internal accuracy of 60% ± 4%. The best mean performance was measured in the SVM algorithm, with a mean internal accuracy of 83% ± 2%. The DT performed almost as well as did the SVM (80% ± 3%), whereas the NNs performed more closely to LR (62% ± 6%).

External validation measures closely matched internal validation measures. The MV data set yielded a mean accuracy of 73% in the validation set, whereas the MV + TRD data set yielded a mean accuracy of 81%. The NN algorithm yielded a mean accuracy of 67% in the validation set, whereas the SVM algorithm yielded a mean accuracy of 87%. The poorest performing
combination was the logistic regression algorithm in combination with MV + TS + CR data set (no TRD features), with an accuracy of 58%. The best performing combination was the SVM algorithm in combination with MV + TRD features, with an accuracy of 94%.

In addition to the ACC measures listed above, SN, SP, and AUROC also were determined for each combination of data set + modeling algorithm. They are provided in Table 1. ROC curves for each data subset of candidate features, using the SVM model, are shown in Figure 4. As was the case for accuracy, the MV + TS data subset had the best performance, with an AUROC = 0.975. Figure 5 shows the complimentary ROC curves contrasting each class of candidate modeling algorithms, using the multivariate + trend feature data subset.

Performing feature reduction with RFE and SVMW on the all-candidate feature data set demonstrated a decrease in mean accuracy from 77% to 75% for RFE and an increase from 77% to 88% for SVMW in the validation data set. Table 2 shows accuracy and AUROC measures models trained on RFE and SVMW feature reduced data subsets compared to all 1025 candidate features.

The final step in analyzing the results was to evaluate the weights assigned to each feature by the various modeling algorithms, looking for conserved patterns across modeling algorithms. Figure 6 shows the plot of the highest ranking features selected by each model class and by both feature reduction algorithms from all 1025 candidate features. No individual features were conserved across all models. Summarizing the distribution across all model classes, MV features accounted for 16% of the features selected, TS features accounted for 46%, CR features accounted for 8%, and TRD features accounted for 30%. The most accurate model (SVM with MV + TRD features) used 51% TRD features and 49% MV features. Figure 7 shows a plot of the top 35 features selected by SVMW, providing the best visual discrimination of arrest cases from
Discussion

This is the first demonstration of a recently proposed clinical prediction modeling strategy that is designed to leverage useful information found in time series data. Almost all clinical prediction models are based on relatively limited multivariate data paradigms in which each feature is assigned a single value. Although these models are adequate for discriminating sick from well patients, they are incapable of discriminating sick and stable patients from sick and deteriorating patients. Thousands of children die in PICUs every year after they experience a cardiac arrest. Many of these arrests are preceded by deteriorations that can be detected as early as 15 to 20 minutes before the arrest occurs. Deteriorations are frequent events in PICUs, and the vast majority does not result in cardiac arrest, so determining which deteriorations are associated with higher risk than others becomes one of clinical intuition. Objective scoring tools (models) that accurately and continuously screen patients for a high risk of arrest can help diminish the cognitive load required to make a judgment of which deteriorations merit treatment and which should simply continue to be monitored. In the same spirit that early warning scores have helped prevent death and disability in acute care units of the hospital, intensive care based warning scores could prevent death and disability in a population that is at much higher baseline risk of death and disability.

This being the first case of use, there are many limitations to the study. First and foremost, although we have demonstrated a significant improvement in baseline model performance by adding TS features to the model, the incidence of cardiac arrest is so low that employing such a model at this stage likely would result in a very high false-alarm rate, quickly dissuading control subjects.
clinicians from relying on it for any real determination of risk. Further refinements are needed in order to boost SN, and more importantly, SP. Specificity for a tool that continuously monitors risk for having a cardiac arrest would need to be very close to 100% in order to be clinically accepted. It happens to be that the SVM model that trained in the MV + TS analysis latent features had a specificity of 100% and a sensitivity of 87%. These numbers are encouraging, but the ratio of controls:events assuming an every-one-minute assessment by the model during the study period, would be in excess of 300,000:1. We have not tested model robustness in this fashion yet because neighboring measures of accuracy suggest that it is highly unlikely to stand up to such scrutiny.

A second limitation to this study is the superficial nature of comparisons. From the candidate feature standpoint, we included only features that relate to shock and were unable to include some of the desired features because they were not available electronically. Also, we included only trend features from the TS elements to establish proof of concept. Numerous other TS analysis measures could be added as additional candidate features, likely further improving the accuracy of the model. From the modeling algorithm standpoint, we included only four candidate algorithms of dozens that are available. Furthermore, we limited our scope to using default model parameters because the number of permutations is quite large. Although default parameters typically work well, they rarely work optimally. A likely example of this limitation can be found in the RFE results. RFE has been shown in many settings to be an excellent method for reducing candidate data sets. However, in our trials, it managed to worsen baseline performance of almost all model classes. As a result, the performance metrics we reported may underestimate the potential of these features to predict cardiac arrest.
We were not surprised to find that the SVM algorithm produced the most accurate models. This modeling algorithm has proven robust in many settings and is touted to work well in both noisy data and in scenarios where case:feature ratios are small. We were somewhat surprised that our attempt to measure the performance effect attributable to the TRD features resulted in the best overall performance measures, even outperforming the features selected by SVMW. We’re not sure if this is a case in which a reasoned strategy outperformed raw statistical measures or if it was just a coincidence. Based on average effects across many modeling runs, using both internal and external validity measures, we believe that the raw TS features tend to serve over-fitting more than discrimination among classes. This effect can be mitigated by using algorithms that tolerate low case:feature ratios (such as support vector machines).

From our perspective, the most meaningful finding in this study is that the TS analysis results that we used as latent features in the model were equally important in discriminating cases from controls as were the MV features. We created 220 candidate trend features, but only 18 were selected. There was no apparent pattern to which features were selected: slopes, intercepts, and averages all were used, as were all raw physiologic variable types. Feature times ranged from 5 minutes to 60 minutes pre arrest. TRD features derived from the clinically relevant features (CR) were not part of the 18 selected features. This finding supports extending the scope of TS analysis to generate other candidate latent features for modeling cardiac arrest in a PICU.

Conclusions

A traditional model using linear regression and MV data misclassified cases 3.7 times more frequently than did the one that used a SVM algorithm in combination with a MV + TRD data subset. Latent features based on combinations of clinical features did not improve model
performance significantly, either in their raw form, or in the TS analysis determinations. On average, acute deteriorations in most vital signs occur within five minutes of an arrest. Drops in pulse oximeter readings and systolic blood pressure begin an average of 15 and 20 minutes before an arrest. We have successfully demonstrated how TS features can be used to improve the predictive accuracy of a clinical prediction model for cardiac arrest in a PICU. Although these findings have significant potential to improve identification of patients at risk for cardiac arrest, and to prevent death and disability by avoiding their occurrence, further refinements are needed to improve model specificity prior to application in a real-world setting.
Figure 1: Raw variables selected to be candidate features for modeling cardiac arrest in a pediatric intensive care unit using time series analysis. Single measurements from the physiologic monitor log that were closest to the arrest event were classified as multivariate features.
Figure 2: Five candidate modeling data subsets used to measure incremental effects of three different classes of candidate features: time series, clinical latent features, and trend analysis latent features. The MV data set consists of the baseline multivariate data. The TS data set includes the multivariate data set + raw time series data. The CR data set includes the TS data set + clinical latent features. The TRD data set includes the multivariate + trend analysis latent features. All features combined measures the net effect of all 1025 candidate features, and is most susceptible to overfitting by virtue of having the lowest case:feature ratio.
Figure 3: Mean values for heart rate (position on y-axis), oxygen saturation (dot shade), and systolic blood pressure (dot size) are shown for arrest (dark) and control (light) subjects over the span of 12 hours. Left to right, the first 12 values represent hourly averages, and the last 60 values represent minute-by-minute measurements. Heart rate indicators occurred very close to the arrest event (acute drop 1-2 minutes beforehand). Pulse oximetry and blood pressure indicators started noticeable trends downward at approximately 20 minutes and 15 minutes beforehand, respectively.
Table 1: Matrix for measures of model performance in validation data set. ACC = accuracy, SN = sensitivity, SP = specificity, ROC = area under the receiver operating characteristic curve. For each performance measure, the highest scoring model class (LR = logistic regression, DT = decision tree, NN = neural network, SVM = support vector machine) is listed to the far right in **boldface** type. Similarly, the highest scoring feature combination (MV = multivariate, TS = multivariate + raw time series, CR = multivariate + time series + clinical latent features, TRD = multivariate + trend analysis latent features, ALL = all 1025 candidate features) is listed in the last row of the performance measure in **boldface** type. The best performing single combination of model class + feature set is listed in **boldface** type in the body of the table for each of the performance measures as well.

<table>
<thead>
<tr>
<th></th>
<th>MV</th>
<th>TS</th>
<th>CR</th>
<th>TRD</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LR</td>
<td>78%</td>
<td>69%</td>
<td>58%</td>
<td>77%</td>
<td>69%</td>
</tr>
<tr>
<td>DT</td>
<td>72%</td>
<td>88%</td>
<td>80%</td>
<td>81%</td>
<td>86%</td>
</tr>
<tr>
<td>NN</td>
<td>67%</td>
<td>63%</td>
<td>67%</td>
<td>72%</td>
<td>66%</td>
</tr>
<tr>
<td>SVM</td>
<td>77%</td>
<td>89%</td>
<td>88%</td>
<td>94%</td>
<td>88%</td>
</tr>
<tr>
<td></td>
<td>73%</td>
<td>77%</td>
<td>73%</td>
<td>81%</td>
<td>77%</td>
</tr>
<tr>
<td>SN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LR</td>
<td>60%</td>
<td>60%</td>
<td>53%</td>
<td>57%</td>
<td>73%</td>
</tr>
<tr>
<td>DT</td>
<td>60%</td>
<td>90%</td>
<td>83%</td>
<td>73%</td>
<td>90%</td>
</tr>
<tr>
<td>NN</td>
<td>33%</td>
<td>20%</td>
<td>30%</td>
<td>40%</td>
<td>27%</td>
</tr>
<tr>
<td>SVM</td>
<td>70%</td>
<td>83%</td>
<td>80%</td>
<td>87%</td>
<td>80%</td>
</tr>
<tr>
<td></td>
<td>56%</td>
<td>63%</td>
<td>62%</td>
<td>64%</td>
<td>68%</td>
</tr>
<tr>
<td>SP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LR</td>
<td>94%</td>
<td>76%</td>
<td>62%</td>
<td>94%</td>
<td>65%</td>
</tr>
<tr>
<td>DT</td>
<td>82%</td>
<td>85%</td>
<td>76%</td>
<td>88%</td>
<td>82%</td>
</tr>
<tr>
<td>NN</td>
<td>97%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>SVM</td>
<td>82%</td>
<td>94%</td>
<td>94%</td>
<td>100%</td>
<td>94%</td>
</tr>
<tr>
<td></td>
<td>89%</td>
<td>89%</td>
<td>83%</td>
<td>96%</td>
<td>85%</td>
</tr>
<tr>
<td>ROC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LR</td>
<td>87%</td>
<td>72%</td>
<td>60%</td>
<td>72%</td>
<td>80%</td>
</tr>
<tr>
<td>DT</td>
<td>74%</td>
<td>90%</td>
<td>79%</td>
<td>83%</td>
<td>88%</td>
</tr>
<tr>
<td>NN</td>
<td>82%</td>
<td>83%</td>
<td>80%</td>
<td>92%</td>
<td>79%</td>
</tr>
<tr>
<td>SVM</td>
<td>82%</td>
<td>95%</td>
<td>96%</td>
<td>98%</td>
<td>97%</td>
</tr>
<tr>
<td></td>
<td>81%</td>
<td>85%</td>
<td>79%</td>
<td>86%</td>
<td>86%</td>
</tr>
</tbody>
</table>
Figure 4: Area under the receiver operating characteristic curve (AUROC) for each class of candidate modeling features (MV = multivariate, TS = multivariate + time series, CR = multivariate + time series + clinical latent features, TRD = multivariate + time series latent features, ALL = all 1025 candidate features) using a support vector machine model (best composite performance modeling algorithm). MV performance was substantially poorer than TRD: 0.820 v. 0.975 area under the curve.
Figure 5: Area under the receiver operating characteristic curve (AUROC) for each class of candidate modeling algorithm (REG=linear regression; DT= decision tree; NN=neural network; SVM=support vector machine) using the multivariate + time series latent features (best composite performance data subset, aka “TRD data set”). REG performance was substantially poorer than SVM: 0.723 v. 0.975 area under the curve.
Figure 6: Highest ranking features plotted on a backdrop of the normalized data set of all subjects and all candidate features. The decision tree algorithm (bottom) limited its representation to only 7 features. 35 features were selected from all other models since 35 was determined to be the optimum number of features that balanced accuracy against over-fitting, based on SVMW ranking. The top set of features represent the best overall performance: support vector machine (SVM algorithm) with only multivariate and trend analysis latent features (TRD data set). All other ranks are based on all 1025 candidate features. No individual features were conserved across all model classes. However, as general classes of features, the best feature reduction algorithm (support vector machine weighting) distributed representations equally across three of the four candidate classes (multivariate = 34%, time series = 32%, and time series latent features = 34%). Clinical latent features were poorly represented across all model classes.
Table 2: Impact of feature reduction on mean model performance, compared to mean model performance in all 1025 candidate features. Recursive feature elimination (RFE) did not have a consistent peak performance, with peaks at both 30 and 40 features (sum of accuracy and area under the ROC curve). Support vector machine weighting (SMVW) peaked performance at 35 features. Recursive feature elimination served to degrade performance while support vector machine weighting substantially improved performance.
Figure 7: Top 35 features determined by support vector machine weighting (weight cutoff = 0.5) in descending order of mean feature intensity of the first 103 subjects. Arrest cases tend to read from left to right: dark to light, whereas control subjects tend to read: light to dark. Boxes approximate the divisions.
References


death with a medical emergency team: Preliminary results. Arch Dis Child 2005;90:1148-
1152

11. Naeem N, Montenegro H: Beyond the intensive care unit: A review of interventions aimed at

support. Crit Care Med 2006;34:2702-2703

mortality and code rates outside the ICU in a children's hospital. JAMA 2007;298:2267-2274


Crit Care 2009;15:203-208

urgent assistance from a medical emergency team. Pediatrics 2008;121:e1577-1584


18. Edwards ED, Powell CVE, Mason BW, et al: Prospective cohort study to test the
predictability of the Cardiff and Vale paediatric early warning system. Arch Dis Child
2009;94:602-606

19. Egdell P, Finlay L, Pedley D: The PAWS score: Validation of an early warning scoring
system for the initial assessment of children in the emergency department. Emerg Med J
2008;25:745 -749


SUMMARY

We have presented three manuscripts as a dissertation of the research we have conducted in using time series analysis to improve clinical predictive modeling. The motivation for this research came from the fact that many children experience cardiac arrests in pediatric intensive care units, and when these cases are reviewed, deteriorations that precede the arrests are often found. The impetus for this research stemmed from the observation that the tools and models commonly employed for prediction in clinical medicine rely on multivariate data structures and do not leverage any information from trends found in data, which essentially defines the ‘deterioration’ that precedes many of these arrests.

As we developed the theoretical framework for encoding ‘deterioration’ into the modeling process, we realized that our proposed strategy had the potential to be applied in other settings where historical observations influence the interpretation of the current state. We also identified a number of new issues that time series data formats introduce into the modeling process. The manuscripts we are presenting detail these issues and provide not only the potential implications and barriers that they present to implementing a model that is based on time series data, but also provides an overview of available strategies that can be used to address the issues and overcome the barriers.

Another broadly applicable result of this research endeavor was the recognition that the trend features found in the time series data elements are but a fraction of the potentially useful candidate modeling features contained in time series data. Although we did not explicitly explore these other features in the cardiac arrest prediction example, we believe they also have a great deal of potential utility in clinical predictive modeling. Both time and frequency domain analysis of seasonality features of time series data is a field that has received attention as single points of focus, but these analyses have not yet been integrated into a larger, multidimensional paradigm.
such as the one we propose. In isolation, these analyses (such as ST-analysis, beat to beat variability / power spectral analysis) have proven to have valuable information, and being able to utilize serial assessments of these features over time is likely to be of particular value in predicting risk of cardiac arrest in both pediatric and adult populations.

The manuscripts we have included in this dissertation have shown utility in performing a conceptually simple time series analysis to improve predictive model accuracy in a clinical setting. Although the procedure of integrating the time series analysis into the modeling paradigm is somewhat more complicated, automation of the key processes is possible, making long term use of these time-series based models feasible. We view the results from the studies we have performed to be pilot data for a much larger corpus of research in the field. To that end, we have identified at least three avenues of study that we intend to carry forward. Two have already had groundwork laid, and the third is in a stage of infancy since it is dependent on the success of the other works.

First, the direct extension of this work is to extend the modeling paradigm to include features from the time and frequency domain analyses of the time series data. Serial measurements of the standard electrocardiogram (ECG) features (P, QRS, T wave analyses and their relationship to one another), beat-to-beat variability, power spectral analysis, and Poincare’ plots are all established methods of analyzing waveform data from an ECG tracing. Inputting the outputs of each of these analyses into a cardiac arrest prediction model will be the next step in this field of research. To that end, I have been granted an AHRQ postdoctoral training fellowship in patient safety and quality to partially support this research (Grant number 1T32-HS017586). Under this training grant, I will continue to have Dr. Jim Turley as my primary mentor. In addition, I have established a new relationship with Dr. Sandra Hanneman, who is the Associate
Dean of Research at the Center for Nursing Research at the University of Texas School of Nursing. The guidance afforded by these two individuals, along with Dr. Virginia Moyer, who is the Chief Quality officer at Texas Children’s Hospital and has committed to serving as a third mentor, will provide fertile soil to help ensure the growth of a successful research career.

One of the expectations of my training in patient safety and quality is to apply for R01 funding. To that end, I have already drafted an R01 application that has undergone two internal reviews. I anticipate submitting this grant before year’s end.

A second avenue of study that has resulted from this research is recognition that the methods likely apply to adults equally well as to pediatrics. I have entered into conversations with a collaborator at The University of California, Los Angeles who has also done some related work in time series analysis for clinical application. Our approaches are somewhat different, but our goals are aligned and there is shared enthusiasm to collaborate on a combined adult / pediatric cardiac arrest prediction project. We have not yet produced a grant application, but we have both committed to doing so before Spring 2011.

Finally, we have demonstrated that time series analysis significantly improves the predictive accuracy of cardiac arrest prediction. However, employing this strategy in a real-time, continuously monitoring paradigm has not been done. We believe it is still too premature to do this, as our extrapolated false positive rate will still be unacceptably high. However, we anticipate that the additions of the waveform analyses and/or the techniques being developed at UCLA are likely to reduce the false positives to a level of clinical acceptance. At that point, we intend to transition our offline, retrospective paradigm to one compatible with real-time application. Although we have identified this as a future endeavor, we have not yet developed the operational details necessary to achieve it.
In conclusion, the path to developing a method for leveraging utility from time series data in making clinical predictions, pediatric cardiac arrest in particular, has been very rewarding from multiple perspectives. First, it has allowed me to establish a number of valued relationships with leaders in the field of health informatics, and to learn from them both the didactic and the practical information that they have shared with me over many years of training. Second, it has provided me with a skill set that is both intellectually and emotionally gratifying, and one that should continue to be so for many years to come. Finally, I believe there is a real possibility that this work can help improve the care that we provide to the children we serve in the pediatric intensive care environment, by helping prevent death and disability that comes from avoidable cardiac arrest scenarios. This is the most rewarding of all, and I intend to push forward with this work until that dream has been realized.

Curtis Kennedy, M.D., M.S.
October 5, 2010