THE DIGITAL PATIENT: MACHINE LEARNING TECHNIQUES FOR ANALYZING ELECTRONIC HEALTH RECORD DATA

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Suchi Saria
August 2011

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Daphne Koller, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Anna Penn

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Sebastian Thrun

Approved for the Stanford University Committee on Graduate Studies.

Patricia J. Gumport, Vice Provost Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.

Abstract

The current unprecedented rate of digitization of longitudinal health data — continuous device monitoring data, laboratory measurements, medication orders, treatment reports, reports of physician assessments — allows visibility into patient health at increasing levels of detail. A clearer lens into this data could help improve decision making both for individual physicians on the front lines of care, and for policy makers setting national direction.

However, this type of data is high-dimensional (an infant with no prior clinical history can have more than 1000 different measurements in the ICU), highly unstructured (the measurements occur irregularly, and different numbers and types of measurements are taken for different patients) and heterogeneous (from ultrasound assessments to lab tests to continuous monitor data). Furthermore, the data is often sparse, systematically not present, and the underlying system is non-stationary. Extracting the full value of the existing data requires novel approaches.

In this thesis, we develop novel methods to show how longitudinal health data contained in Electronic Health Records (EHRs) can be harnessed for making novel clinical discoveries. For this, one requires access to patient outcome data — which patient has which complications. We present a method for automated extraction of patient outcomes from EHR data; our method shows how natural languages cues from the physicians notes can be combined with clinical events that occur during a patient's length of stay in the hospital to extract significantly higher quality annotations than previous state-of-the-art systems.

We develop novel methods for exploratory analysis and structure discovery in bedside monitor data. This data forms the bulk of the data collected on any patient yet, it is not utilized in any substantive way post collection. We present methods to discover recurring *shape* and *dynamic* signatures in this data. While we primarily focus on clinical time series, our methods also generalize to other continuous-valued time series data.

Our analysis of the bedside monitor data led us to a novel use of this data for risk prediction in infants. Using features automatically extracted from physiologic signals collected in the first 3 hours of life, we develop Physiscore, a tool that predicts infants at risk for major complications downstream. Physiscore is both fully automated and significantly more accurate than the current standard of care. It can be used for resource optimization within a NICU, managing infant transport to a higher level of care and parental counseling. Overall, this thesis illustrates how the use of machine learning for analyzing these large scale digital patient data repositories can yield new clinical discoveries and potentially useful tools for improving patient care.

Acknowledgements

I am greatly indebted to the many people who have contributed to the creation of this thesis.

I am profoundly grateful to my advisor, Daphne Koller. She has been an amazing mentor and a voice of reason throughout. Her tireless pursuit of excellence in everything she undertakes — teaching, mentoring, research, writing, and other aspects of her academic work — is truly inspirational. From her I have learnt a great many skills: how to approach a complex problem; how to critically analyze results; how to push ideas to their full development; how to present research in an engaging manner; and the list goes on. I feel extremely fortunate to have had her as a mentor and will miss being able to walk down the hallway to have candid conversations about work and research.

I would also like to thank my committee members, Anna Penn and Sebastian Thrun, as well as my defense committee members, Jeffrey Gould and Eric Horvitz. Each of them have immensely influenced my way of thinking and this thesis as a result. In particular, Annie provided invaluable clinical guidance on this project. Through her, I was introduced to this direction of research and, without her, this project would not have been feasible. Sebastian's research has continually inspired me and it strongly influenced my choice in tackling this challenging and important real-world domain. Jeff encouraged me with exciting ideas, visionary insights and moral support throughout this work. Eric has been a tremendous mentor throughout. I have learnt valuable things from him, both on the academic and the personal front.

During the last two years, I have interacted closely with Dr. Anand Rajani, a neonatology fellow at the Lucile Packard Children's Hospital. Anand tirelessly answered my questions about neonatology and entertained my speculative conversations about how one might attack this data. In the process, I also acquired a dear friend with whom working on this project became even more enjoyable.

The data collection from the Stanford NICU was the product of generous effort from several different people. Eric Halfenbein, a researcher at Philips, came to our rescue while we were struggling to set up a system to extract monitor data. Eric helped us both setup the system and has helped maintain it in the last three years. Nestor Llerena was key in us having access to the data in the EHR. Purna Prasad and George Yang helped coordinate the monitor data storage and collection effort. Research nurses Patty Hartsell, Judith Hall and Betsy Kogut patiently annotated the NICU data. Judy also helped me with our IRBs.

I feel fortunate to have been part of DAGS and to have shared the company, the ideas, the questions and the expertise of Uri Nodelman, Gal Elidan, Gal Chechik, Suin Lee, David Vickrey, Benjamin Packer, Alexis Battle, Geremy Heitz, Stephen Gould, Jonathan Lasserson, Vladimir Jojic, Pawan Mudigonda, Cristina Pop, Tianshi Gao, Huayan Wang, Varun Ganapathi, Karen Sachs and Sara Mostafavi. In particular, I want to thank Uri, David and Vladimir, my "go to" people over the years, with whom I have had several insightful conversations about research. After many of these conversations, I often remember leaving elated about having learnt something new. Ben, Karen and Sara each helped me dramatically improve and refine the presentation of the ideas in this thesis. Gal Elidan was a wonderful and entertaining sounding board about life-decisions through the years.

During my Ph.D., I have had a chance to mentor and work closely with several students including Andrew Duchi, Laney Kuenzel, Rohini Rajaraman, and Shui Hu. I want to thank you for choosing to work with me and for the insights we developed as a result of the projects. In particular, I would like to thank Andrew Duchi. I greatly enjoyed my time working with him on CSTMs.

I also want to thank Hendrik Dahlkamp and Cristian Plagemann who helped collect the Kinect^{TM} dataset for our experiments in chapter 4.

Many thanks to my friends who have had nothing to do with work in this thesis, but worked hard to keep my relative sanity throughout. I will not list all of you here, but my gratitude to you is immense.

I am at a loss for words to thank my family. I was fortunate to be born in a large and loving extended family. This made for countless childhood memories and a large support network I could always rely on. In particular, my parents, Saroj and Shiv Saria, have sacrificed much to give me the chance to pursue the opportunities available to me. They have given me unconditional love, support, encouragement and everything that a child could ask for including the opportunity to study abroad. To my brother, who is both very talented

and marvelously entertaining. He has kept me on my toes always. And, to John, my tireless partner in work and play, tears and laughter, move fires and travel adventures, thank you for making me so happy next to you.

To my parents — Saroj Saria and Shiv Saria.

Contents

Abstract **v**

Acknowledgements **vii**

1 Introduction **1**

1.1 Electronic Health Record data 2

1.1.1 Our NICU dataset 5

1.2 Contributions per chapter 6

1.3 Challenges in analyzing Electronic Health Record data 7

1.4 Related Work 9

2 Extracting Patient Outcomes **13**

2.1 Introduction 13

2.2 Automated Outcome Extraction 18

2.3 Methods 18

2.3.1 Language Features 20

2.3.2 Clinical features 21

2.3.3 Learning Technique 22

2.4 Experiments and Results 24

2.4.1 Statistical Methods 24

2.4.2 Baseline Language Model 24

2.4.3 Integrated EHR Model 25

2.4.4 Error Analysis 25

2.4.5 Improving human annotation 28

2.5 Discussion and Conclusion 29

3 Discovering Dynamic Signatures in Physiologic data **31**
3.1 Introduction . . . 31
3.2 Background . . . 34
3.2.1 Dirichlet Processes . . . 35
3.2.2 Hierarchical Dirichlet Processes . . . 36
3.3 Time Series Topic Model . . . 38
3.3.1 Overview of the model variables . . . 39
3.3.2 Generative Process . . . 40
3.4 Related Work . . . 42
3.5 Approximate Inference . . . 43
3.6 Experiments and Results . . . 47
3.6.1 Experimental Setup . . . 47
3.6.2 Quantitative Evaluation . . . 48
3.6.3 Qualitative Evaluation . . . 55
3.7 Discussion and Future work . . . 57

4 Discovering Shape Signatures in Physiologic data **59**
4.1 Introduction . . . 60
4.2 Generative Model . . . 61
4.2.1 Canonical Shape Templates (CST) . . . 62
4.2.2 CST Deformation Model . . . 64
4.2.3 Non-repeating Random Walk (NRW) . . . 65
4.2.4 Template Transitions . . . 66
4.2.5 Summary of CSTM generative process . . . 66
4.3 Learning the model . . . 67
4.3.1 E-step . . . 67
4.3.2 M-step . . . 68
4.3.3 Escaping local maxima . . . 72
4.3.4 Model Initialization . . . 72
4.3.5 Preprocessing . . . 74
4.3.6 Summary of learning algorithm . . . 74
4.4 Experiments and Results . . . 74
4.4.1 Baseline Methods . . . 75

4.4.2 Metric . . . 76
4.4.3 Datasets . . . 76
4.4.4 Results . . . 78
4.5 Discussion and Conclusion . . . 84

5 Clinical Application to Risk Stratification 87
5.1 Introduction . . . 88
5.2 Risk Stratification . . . 89
5.3 Methods . . . 89
5.3.1 Feature construction . . . 90
5.3.2 Physiologic signal processing . . . 90
5.3.3 Combining risk factors into a score . . . 91
5.3.4 Learning the score parameters . . . 92
5.3.5 Nonlinear models of risk factors . . . 92
5.3.6 PhysiScore: Probabilistic score for illness severity . . . 94
5.4 Experiments and Results . . . 94
5.4.1 Outcome annotation . . . 95
5.4.2 Study population . . . 95
5.4.3 Statistical methods . . . 96
5.4.4 Results . . . 96
5.4.5 Importance of physiological features . . . 97
5.5 Discussion . . . 100
5.5.1 Discriminative capacity . . . 100
5.5.2 Technical considerations . . . 103
5.5.3 Advanced computational techniques in modern medical settings . . . 103

6 Conclusion 107

Bibliography 111

List of Tables

2.1 List of complication-specific clinical features used. Complications are listed in order of decreasing frequency in our data set. Features are extracted from medications (M), clinical events (E), culture reports (C) and radiology reports (R). Overall, 33 clinical features are extracted. 19
2.2 Language transfer features. 23
2.3 Baseline: language model performance. 25
2.4 Performance comparison between the language model and the EHR model. For visual clarity, the winning model, chosen based on F1, is underlined for each complication. For the outcome labels of death, penumonia, pneumothorax, pulmonary hemorrhage, and pulmonary hypertension, no clinical features were available. 26

3.1 Notation. 39
3.2 Evaluating features from unsupervised training of TSTM. 52

4.1 Notation for the generative process of CSTM and other frequently used notation in this chapter. 62
4.2 Summary of the learning algorithm for CSTM 74

5.1 Baseline and disease characteristics of the study cohort. (SGA, small for gestational age; NOS, not otherwise specified.) 105
5.2 Performance summary with AUCs. 106

List of Figures

1.1 Electronic health record (EHR) data collected from a neonatal intensive care unit (NICU): a) continuous physiologic monitoring data, b) laboratory orders and measurements, c) medications and procedures administered, d) imaging results, e) admission, progress and discharge summary notes. 3

3.1 Heart signal (mean removed) from three infants in their first few hours of life. 32

3.2 Example of three time series with shared signatures. Segments of each distinct color are generated by the same function, and thereby are instances of the same signature or *word*. Signatures here correspond to autoregressive functions. The choice of function used at any given time depends on the latent *topic* at that time. While the three series differ greatly in their composition, they contain shared structure to varying extents. 33

3.3 Graphical representation of the Time Series Topic Model (TSTM). 38

3.4 Experimental protocol for the evaluation of goodness-of-fit, a) the procedure for splitting each series into the train and test set, b) the pipeline for evaluating goodness of fit on the data. 49

3.5 Test log-likelihood from three separate Gibbs chains for the AR(1)-HMM, AR(2)-HMM, and TSTM with an AR(1) observation model evaluated on a) the heart rate data (top), b) the respiratory rate data (bottom). 50

3.6 a) & c) Inferred word distributions from the heart rate signal for 30 infants during their first four days at the NICU with the BP-AR-HMM for two different initializations (initialization setting described in the text); distinct colors correspond to distinct words, b)& d) Corresponding data log-likelihood of the Gibbs chain for the first 5000 iterations. 53

3.7 (a) Inferred word distributions for the heart rate data for 30 infants during their stay at the NICU. At the bottom of the word panel, infants marked with red squares have no complications, (b) distribution over disease topic given words for the population, (c) posterior over latent state, *Healthy*, (d) examples of inferred features extracted from the data. 58

4.1 a) The template S shows the canonical shape for the pen-tip velocity along the x-dimension and a piecewise Bézier fit to the signal. The generation of two different transformed versions of the template are shown; for simplicity, we assume only a temporal warp is used and ω tracks the warp at each time, b) The resulting character 'w' generated by integrating velocities along both the x and y dimension. 63

4.2 Hand-drawn curves from the simulated dataset. a) Example sequences generated from each of the motifs, b) Examples sequences aligned to their canonical shape template. The bottom right subplot in (a) shows that sequences from the NRW state do not have the same repeating structure as those from the CSTs do. The bottom right suplot from (b) shows the data once aligned to the NRW state. Each segment is of unit length once aligned. 78

4.3 Comparison of the CSTM with an initialization using the peak-based method, and initializations from GreedyRPM and GreedyRPC with different settings for d on the character dataset. In the figure, GreedyRPM and GreedyRPC have been abbreviated as Md and Cd respectively. 79

4.4 Accuracy on Character (top) and Kinect (bottom) for CSTM and its variants. Two different initializations for CSTM are compared: GreedyRPM10 and peak-based. 81

4.5 Confusion matrix showing performance of CSTM (left) and CSTM-DTW (right) on the character data. 82

4.6 Classification performance for increasing level's of NRW to CST proportion in the data. 83

4.7 a) An example bradycardia cluster extracted by GreedyRPM, b) an example bradycardia cluster recovered by CSTM, c) the bradycardia cluster shown with sequences aligned to the shape template. Note that the bradycardia sequences extracted by CSTM are more heterogeneous in appearance than those captured by GreedyRPM. Thus, CSTM is able to better capture the variability in the cluster. 83

4.8 ROC curve for recovering bradycardia sequences from the data using CSTM and GreedyRPM with various parameter settings for both models. 85

4.9 Two examples of novel clinical events, a) mutliple varying level's of bradycardia like episodes (bradycardia in a preemie is defined as HR $<$ 80 for longer than 10 seconds) within a short span of 120 seconds, b) fast and large oscillations in heart rate. 86

5.1 Processing signal subcomponents. Differing heart rate variability in two neonates matched for gestational age (29 weeks) and weight (1.15 $\pm$ 0.5 kg). Original and base signals are used to compute the residual signal. Differences in variability can be appreciated between the neonate predicted to have HM (right) versus LM (left) by PhysiScore. 91

5.2 Distribution of residual heart rate variability (HRvarS) in all infants. Learned parametric distributions overlaid on the data distributions for HRvarS displayed for the HM versus LM categorization. 93

5.3 (A) ROC curves demonstrating PhysiScores performance in predicting high morbidity as it relates to conventional scoring systems. (B) PhysiScores performance with laboratory studies. (C) Predictions for infants with infection-related complications. (D) Predictions for infants with major cardiopulmonary complications. 98

5.4 ROC curve demonstrating the limited sensitivity of PhysiScore in predicting morbidity for infants with IVH. Each circle represents the IVH grade of a preterm neonate overlaid on the respective score. 99

5.5 The significance of different physiological parameters in predicting high morbidity. (A) The learned weight (w_i in Eq. 5.1) for each physiological parameter incorporated in PhysiScore; error bars indicate variation in the weight over the different folds of the cross-validation. (B) The nonlinear function associating the parameter with the risk of high versus low morbidity. 101

Chapter 1

Introduction

Our interactions with the health care system are getting digitized at a rapidly accelerated pace. In the majority case, recordings of this interaction include early presentation of the symptoms, the sets of diagnostic tests administered and their results, passive monitoring results, the series of interventions, and detailed reports of health progression by the health practitioner. In some cases, these recordings can be as detailed as the inclusion of video data from the physician-patient interaction [Field and Grigsby, 2002].

This high-granularity recording of the actual patient, which can be stored, shared and retrospectively analyzed creates the *digital patient*. With machine learning and data analysis methods, we can analyze the digital patient population to retrospectively make inferences about our health care system. For example, reasoning with information extracted from data across multiple visits across a large population can illuminate the evolution of a given disease and the evolution of an individual's health status as compared with the population. Comparing patient health trajectories across individual physicians, physician practices and even hospitals can provide a way to compare effectiveness across physicians, practices and hospitals. This can be used to infer intervention protocols that lead to better health trajectories and thereafter, guide clinical practice guidelines or inform national policy decisions. At the point-of-care, knowing relevant facts from the patient's clinical history can significantly improve patient-physician interaction and help avoid medical errors (e.g., missed drug-drug interactions). A retrospective analysis of the data can potentially identify subpopulations that respond more effectively to interventions. This data can also lead to the discovery of new markers for early prediction of disease.

We describe in detail below the different types of data collected in the Electronic Health

Record (EHR). Information required to answer any of the problems posed above cannot be simply *extracted* from the EHR database; most of the required information is not directly observed and must be *inferred* from this data. This data is high-dimensional, heterogeneous, highly unstructured, large-scale and noisy, which makes the task of modeling and inference from this data challenging. In this thesis, we seek to address the tasks of clinical discovery and prediction from EHR data via novel probabilistic methods to model the data.

1.1 Electronic Health Record data

Electronic Health Record databases contain patient encounter data recorded to varying levels of granularity. Increasingly, due to the Meaningful Use legislation [111th United States Congress, 2009a], hospitals and physicians are capturing increasing amounts of data at the highest levels of granularity.

Below, we describe an example dataset from Stanford's Lucile Packard Children's Hospital (LPCH). The data is collected from infants in their Neonatal Intensive Care Unit (NICU). See figure 1.1 for an overview of the different types of data.

- **Continuous physiologic monitoring data:** Bedside monitors are used extensively in most intensive care units for continuous monitoring of physiologic data. At the Stanford NICU, heart rate (HR), respiratory rate (RR) and oxygen saturation (OS) data is captured on all infants starting at the time of admission (see figure 1.1a for example recordings of HR, RR and OS data). Other signals of mean, systolic and diastolic blood pressure, ectopic counts (count of abnormal heart beats), and pulse waveforms are collected at the physician's discretion. Noise sources such as lead drops, handling of the infant and feeding can corrupt the data in spurts. Data is missing when the infant is away for treatment or during transport.

- **Laboratory measurements:** A wide variety of laboratory tests are performed on these infants. Several of these measurements (e.g., the complete blood count (CBC)) are started soon after birth and performed repeatedly, often at fixed intervals. Other measurements are initiated to be recorded repeatedly or recorded once as needed basis. The EHR contains both time when the test was ordered and the time of the actual measurement along with the result of the measurement. The nurses typically record the measurement as its taken. Occassionally, there can be a delay between when the

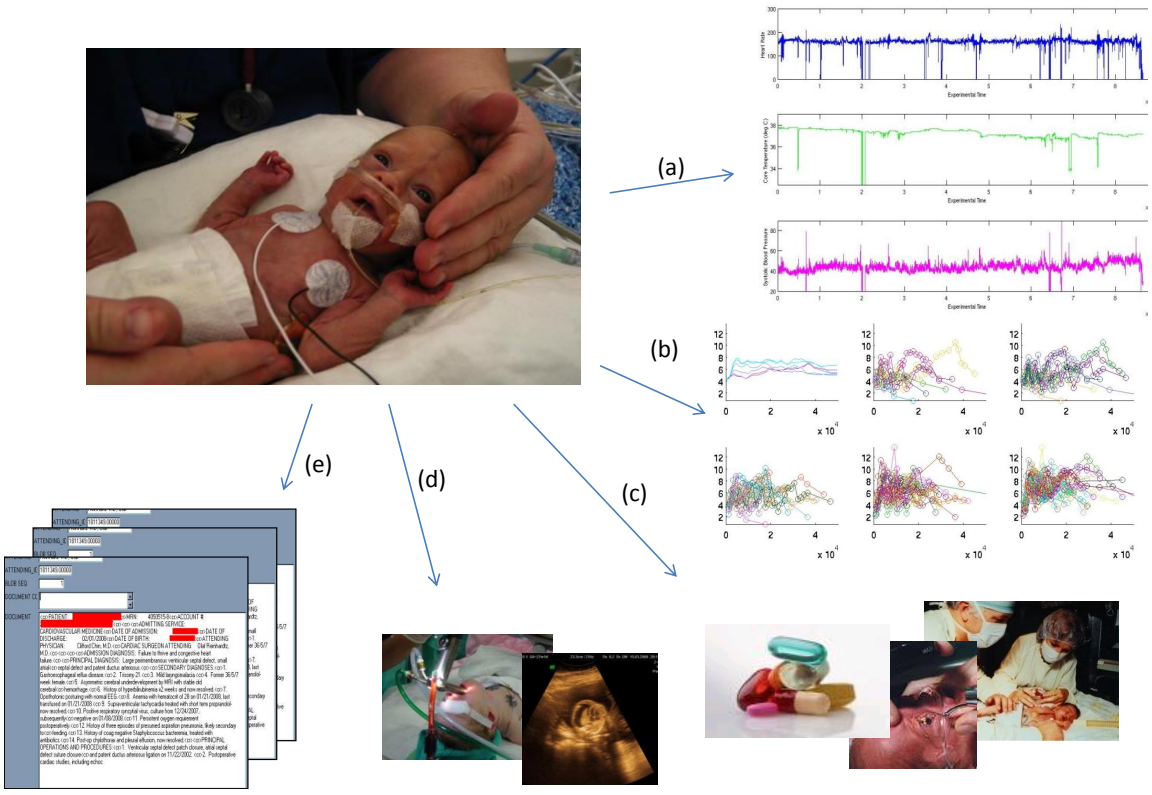


Figure 1.1: Electronic health record (EHR) data collected from a neonatal intensive care unit (NICU): a) continuous physiologic monitoring data, b) laboratory orders and measurements, c) medications and procedures administered, d) imaging results, e) admission, progress and discharge summary notes.

measurement was taken and when it was recorded. This delay is not recorded in the data, producing a source of noise. For the measurements that are often frequently recorded (e.g., vital signs), if the measurement is very different from what is expected, sometimes the nurses avoid recording it or modify it to reflect their notion of what they think is going on. Other sources of noise include the test's own measurement error which varies from test to test.

- **Medications administered:** For every medication, the time, dosage and frequency of medication is recorded at the time that it is first prescribed. In addition, each time

the medication is administered, a record of the dosage and time since birth is made.[1] Below, we show an example of records for a randomly selected infant.

PatientID — Time since birth — Name — Dosage — Units
530044567—14.00:00:00—Sodium Chloride 0.45%—0.999001—mL
530044567—14.00:00:00—heparin—0.999001—units
530044567—14.00:00:00—IVPARENT—1.000000—mL
530044567—14.01:00:00—fat emulsion, intravenous—0.300000—mL
530044567—14.01:00:00—IVPARENT—0.300000—mL
530044567—14.02:00:00—fat emulsion, intravenous—0.300000—mL
530044567—14.02:00:00—IVPARENT—0.300000—mL
530044567—14.02:00:00—furosemide—0.600000—mg
530044567—14.03:00:00—parenteral nutrition solution—2.000000—mL
530044567—14.03:00:00—IVPARENT—2.000000—mL
530044567—14.03:00:00—Sodium Chloride 0.45%—0.999001—mL
530044567—14.03:00:00—heparin—0.999001—units
530044567—14.03:00:00—IVPARENT—1.000000—mL

- **Treatments and Procedures:** The time and names of all treatments and procedures prescribed are recorded as events. In addition, the location and whether or not anesthesia was given is also recorded. The latter two data elements are critical in reasoning about noise or missing data. Below, we show example events. Time since birth of 00 : 00 : 00 denotes that the corresponding event took place soon after birth but recorded to be at birth.

PatientID— Time since birth — ProcedureID — NomenclatureID — Priority — Anesthsia — Anesthesia minutes — Location — Procedure name
531248861—00:00:00—39157910—1275040—1—None—0—2W—Insertion of endotracheal tube
531248861—00:00:00—40533850—1276338—3—None—0—2W—Injection or infusion of other therapeutic or prophylactic substance
531248861—00:00:00—40533853—1258053—5—None—0—2W—Umbilical vein

[1]The data is deidentified to record time as <day><day> . <hour><hour>:<minute><minute>:<second><second> and generate random patient IDs.

catheterization
531248861—00:00:00—40533855—1275188—6—None—0—2W—Gastric gavage
531248861—00:00:00—40533857—12280077—7—None—0—2W—Non-invasive mechanical ventilation
531248861—00:00:00—40533859—1250601—8—None—0—2W—Spinal tap
531248861—4.10:00:00—39157908—1258057—9—Local—0—2W—Venous catheterization, not elsewhere classified
531248861—6.00:00:00—40533861—1276578—10—None—0—2W—Other phototherapy

- **Imaging results:** X-rays and MRI scans are recorded either as part of routine care or on a need basis by the physician. The EHR contains the original image data and the analysis of the image made by the treating physician and/or radiologist.

- **Physician Notes:** The bulk of the data (in terms of size) are notes by the care provider. Typically, an admission report detailing symptoms and conditions at the time of admission is made. Progress notes are recorded at the time of any major events and on a daily basis. At the time of transfer to a different unit or discharge from the hospital, a discharge summary is dictated that includes a detailed summary of most major events during the entire length of stay in the NICU. An infant at the NICU for a period of a month may be attended to by more than 20 doctors, 40 nurses, and scores of other care takers. The final discharge summary is often written after reviewing previous records and from mental recollection. These notes are used as hand-off tools between caretakers and for the purpose of billing so effort is made to make these notes comprehensive. We tackle the task of extracting patient outcomes from these notes and show a detailed example in context later in chapter 2.

1.1.1 Our NICU dataset

For this thesis, we use data from the LPCH NICU as an example EHR dataset from which we motivate or demonstrate several of the ideas presented in later chapters. Thus, we briefly describe the specifics of our dataset here.

Our data collection process was initiated in March 2008 and has been ongoing since. We have been capturing data from all infants admitted to the NICU after March 2008 and who are within their first week of birth. For our studies, we only use data from premature infants with gestational age ≤ 34 weeks and weight ≤ 2000 grams. These infants are more likely to have problems associated with prematurity. Approximately, 150 infants have met this criteria per year, resulting in approximately 100Gb of data collected per year.

Stanford collects monitor data via the Philips bedside monitor systems. Data is collected at a sub-second granularity, but stored for retrospective analysis only at the minute-level granularity. While storing data at higher granularities was a possibility, the storage costs are much higher. Therefore, the tradeoff for storing data at the minute-level granularity was made. This is typical at other institutions.

Our study is covered under the Stanford IRB protocol 8312. Each of our studies, presented in later chapters, had additional inclusion criteria based on which smaller subsets of this dataset was used. We describe the population characteristics for each subset in the chapters when they are used.

1.2 Contributions per chapter

This thesis is a foray into how observational patient data from the EHR can be harnessed for making novel clinical discoveries. While the data is extremely rich, it is also extremely challenging and requires developing a deep understanding of the domain and the data collection process, learning about relevant clinical biases to guide model building, and novel methods to surmount dimensionality and noise related issues in the data.

In section 1.3 below, I describe some of the common challenges that machine learning practitioners must face in handling this data. One such challenge is the lack of gold standard patient outcome data (which infant has which complications); this data is necessary for measuring performance on most clinically relevant prediction tasks including that of risk prediction which we tackle later in this thesis. In chapter 2, we describe a method for extracting patient outcomes from the EHR data. Previous methods have performed this task primarily from free-text data contained in the discharge summaries. In addition, we exploit structured data in the EHR to significantly improve performance.

Bedside monitor data is continuously collected on all infants from the time of admission until discharge and forms the bulk of the data collected on any patient. However, this

data is not used offline post collection in any substantive way by the care givers. In our method, its use online is limited to answering simple questions such as whether the heart rate is too high or the respiratory rate is too low. In chapters 3 and 4, we develop unsupervised methods for discovering structure and repeating signatures (e.g., dynamics and shape signatures) from this continuous monitor data, and relating these signature patterns to diseases. These methods incoporate clinical biases relevant to the data into probabilistic models. This improves performance significantly compared with previous time series processing methods. We evaluate performance both qualitatively, to assess consistency with known clinical results, and quantitatively.

Our analysis of the bedside monitor data led us to novel use of this data for risk prediction. In chapter 5, we build on this insight to develop and validate a tool, akin to the electronic Apgar [Casey *et al.*, 2001] (a score universally used to assess an infant's health at birth), for early prediction of morbidity in premature infants. Our tool is significantly more accurate than Apgar and it can be fully automated. We also compare with other previously developed neonatal monitoring scores that require invasive tests and show that our method outperforms those using just routinely collected non-invasive data. Overall, our thesis illustrates how the use of machine learning for analyzing digital patient repository data can yield new clinical discoveries and potentially useful tools for improving patient care.

1.3 Challenges in analyzing Electronic Health Record data

There are several challenges that dominate the use of electronic health record (EHR) data. While addressing these challenges is necessary to successfully tackle any of the above mentioned health care questions, these challenges also provide interesting computational problems.

First, obtaining high-quality data is essential, and its lack can hurt the significance of the conclusions that can be drawn from the data. While this is often the case in most domains, this is especially true when working with observational data such as patient health record data. For example, patient outcomes are typically stored using billing codes called ICD9 codes. However, these codes are known to have poor granularity and accuracy for identifying patient outcomes [Campbell and Payne, 1994; Solti *et al.*, 2008]. Furthermore, since those are coded for billing, often complications that cannot be billed are not noted.

Or, complications are often assigned codes that help optimize reimbursement. Therefore, the data as is does not provide accurate outcome information against which performance can be reliably measured. In another example, at Stanford, the bedside monitoring device flushes physiologic signal information to a central repository every $6-8$ hours or each time a bed change occurs. Each signal file is appended with identifiable information such as name, date of birth and medical record number of the infant. The device requires manual entry of this information by the nurses each time an infant is moved from one bed to another. On multiple occasions, the infant will have been transferred but his information was not updated on the monitor. As a result, the record from the infant who was previously on the same bed would get erroneously attached to the physiologic data from the current occupant while the current occupant's data would appear to be missing. At other times, a record of the move was successfully logged on the new monitor but the infant was not transferred out of his previous bed. Thus, it would appear as if the infant was on two beds at the same time. Several of these errors are corrected by inferring the bed status of the infant based on movement through the system and analyzing the contents of the file to ensure consistency across file boundaries. However, the use of the data in its original form without realization of this measurement bias will lead to inaccurate conclusions. Therefore, understanding these biases in data collection are important in obtaining robust and high quality data.

Second, the data is aggregated from several different modalities. A thousand unique laboratory tests are done in the neonatal ICU alone. Most of these tests have different noise characteristics. For example, the neutrophil (a type of white blood cell) measurements are performed by manually counting the number of cells of each neutrophil type under a microscope using a clicker. Moreover these cells are described by the technician and it is a very subjective process. Thus, the observations may include error that is technician dependent and varies widely between shifts. Similarly while monitoring continuous blood pressure, measurements may be corrupted by a variety of reasons: when a blood sample is drawn, the sensor drops, the patient is handled, or the infant in the neighboring crib is crying [Aleks *et al.*, 2009; Quinn *et al.*, 2009]. While a model-driven approach that models each sensor modality in detail is useful, modeling all sources of variability and noise may not always be feasible. Therefore, progress must be made cognizant of this shortcoming.

Third, the data is high-dimensional. A supervised training approach of using labeled data to select the relevant dimensions in this data is often infeasible; annotated data is scarce because collecting patient data is expensive and requires input from skilled practitioners or

specially trained data recorders, such as research nurses. Dimensionality reduction methods that can extract informative representations from the data can both reduce the need for labeled data, and yield more meaningful features for prediction tasks. For example, for the data measured by the monitor continuously, features representing specific shapes (a repeated drop and a rise in the heart rate) are known to be clinically relevant. The discovery of these shapes and inclusion of higher-level features such as the frequency of individual shapes can be more informative than the raw signal data itself. Similarly, a bag of words model that includes all words from the clinical narratives (progress notes or discharge summaries) may be less effective than higher-level features that are a result of parsing. Methods for dimensionality reduction that can incorporate clinical biases to inform the reduction of this data are imperative.

Another challenge is the lack of metrics. Most problems lack existing baselines and formal metrics against which they can be easily calibrated. This slows down progress. For example, in assessing the efficacy of unsupervised discovery techniques, usefulness of the extracted features must be established against supervised tasks where other simple features do not suffice. Additionally, when possible, a qualitative validation of consistency with known previous clinical results provides another means of establishing confidence in the results.

1.4 Related Work

Data contained in the EHR has been worked on by various communities for more than a decade. While a comprehensive review is not feasible, work closely related to each of the different aspects (underlined in the text) of this thesis is discussed below. Individual chapters discuss additional works related primarily to the data or model described in that chapter.

The Physionet project [Goldberger *et al.*, 2000], a collaborative effort between MIT, Harvard's Beth Israel Hospital and Boston University created the first publicly available EHR data repository for adult ICU data. They created a database called MIMIC that contains comprehensive EHR data from a single ICU visit. However, the identity of a patient is not tracked across multiple visits; thus, data across multiple visits for any single patient cannot be tracked. Most adult ICU patient visits last from a few hours to a couple of days so only short term monitoring of health status is feasible using this data.

The MIMIC database has spawned a body of work on physiologic signal processing from high granularity beat-to-beat heart rate data. Detrended Fluctuation Analysis [Peng *et al.*, 1995], is a method for studying the fractal characteristics of beat-to-beat variability in the heart rate, measures correlations in the beat-to-beat signal over short- and long-ranges. Altered correlation patterns have been found in signals from both patients with amyotrophic lateral schlerosis [Hausdorff *et al.*, 2000], congestive heart failure [Poon and Merrill, 1997] and infants with intraventricular hemorrhage [Tuzcu *et al.*, 2009]. Syed et. al. [2009a] have developed a method where beat-to-beat heart rate signals are windowed and similar windows are found using random projections. A frequency spectrogram of the resulting windows were found to be predictive of morbidity in patients with acute coronary syndrome within 90 days of a non-ST-elevation [Syed *et al.*, 2009b]. A large amount of work has been done in the signal processing community related to electrocardiogram (EKG) data (see [Moody and Mark, 2001] for examples). Overall, most approaches are not general purpose and tailored to a specific signal.

Techniques for constructing abstractions or lower-dimensional representations of multiparameter clinical time series data for the purpose of hypothesis discovery have been developed extensively. These methods typically manually identify characteristics of intervals that are clinically relevant for individual diseases. Alternately, they extract general purpose symbols (e.g., decrease, increase or constant) which are further abstracted to form strings of symbols to represent complex phenomena in the time series. These methods are work-intensive, and rely on clinical experts and heavy manual knowledge engineering (see [Stacey and McGregor, 2007] for a survey).

The majority of work on bedside monitoring has focused on generating alarms. Known clinically-relevant signatures (e.g., is the heart rate at the current time beyond the clinical norm or is the patient undergoing apnea [Mietus *et al.*, 2000]) are detected in real-time via online analysis of the continuously streaming measurements. Another line of work has focused on reducing false alarm rates by modeling false alarm events such as sensor drops or a blood sample draw [Quinn *et al.*, 2009]. In both cases, the goal is of automation and not discovery.

Both in the neonatal and adult population, risk prediction tools have been developed for measuring severity of disease. For example, in adults, the APACHE and APACHE II scores [Knaus *et al.*, 1985] combine measurements such as temperature, heart rate, blood pressure and so on to make assessments in the first 24 hours about illness severity for patients in

the ICU. Neonatal risk prediction scores of CRIB [Network, 1993], SNAP and SNAPPE [Richardson *et al.*, 2001] combine laboratory measurements and vital signs at 12 hours to quantify overall health. All of these scores require invasive tests and manual intervention. Outside the ICU, numerous other works have used clinical data and Bayesian networks for building expert systems for clinical decision support (e.g., [D. Heckerman and Nathwani, 1992]). More recently, clinical events from the EHR have been combined for predicting individual conditions (e.g., [Reis *et al.*, 2009; Himes *et al.*, 2009]).

A different community of researchers has focused on using natural language process for information extraction from free-text data contained in the EHR. A popular approach here is to use a concept indexing system which seeks to map text to standardized concepts using terminologies such as those in the Unified Medical Language System (UMLS). The identified concepts are then used for a variety of extraction tasks [Meystre *et al.*, 2008]. For example, generating medication lists and patient problem lists, extracting coded data for decision support systems, or automatic detection of adverse events [Melton and Hripcsak, 2005].

Recent efforts have been made to build large databases that consolidate data across multiple institutions for clinical discovery especially in relatively infrequent and complicated diseases [Liao *et al.*, 2010]. Most studies on discovery are based on association rules between terms discovered in the free-text (e.g., [Goldacre *et al.*, 2009; Crawshaw *et al.*, 2010; Petri *et al.*, 2010]). These methods cannot be easily extended to incorporate domain knowledge or various types of clinical biases.

For specific application of medical diagnosis, a large amount of work has been pursued in the machine learning and informatics community.

Novel flexible methods that can discover structure while modeling both clinical and measurement biases can lead to potentially new and powerful medical discoveries.

Chapter 2

Extracting Patient Outcomes

Access to patient outcomes is necessary to leverage EHR data within any machine learning system. Traditionally, outcomes are manually coded by physicians or trained coders. More recently, natural language processing based systems have been used to automatically extract outcomes from clinical free-text data such as the physician's dictation. Integrating easy-to-extract structured information such as medication and treatments (available in the EHR) into current natural language processing based systems can significantly boost coding performance; in this chapter, we present a system that rigorously attempts to validate this intuitive idea. Based on recent i2b2 challenge winners, we derive a strong language model baseline that extracts patient outcomes from discharge summaries. Upon incorporating additional clinical cues into this language model, we see a significant boost in performance to F1 of 88.3 and a corresponding reduction in error of 23.52%. Moreover, to obtain gold-standard data from the nurses, we used our system to corroborate their annotations iteratively, and we found significant reductions in both coding time required and the fraction of human errors reduced.

2.1 Introduction

The modern hospital generates large volumes of data, which include discharge summaries, records of medicines administered, laboratory results and treatments provided. With the recent ubiquity of electronic medical record (EHR) databases, all of this patient information is often documented within a single storage system. However, one of the key difficulties in using this EHR data is that patient outcomes – who has what complications – the most

basic information required to assess performance of any learning algorithm for this data is not readily available. At present, in most systems, patient outcomes are manually coded by the physician for billing purposes primarily. As a result, the codes are not comprehensive [Campbell and Payne, 1994; Solti *et al.*, 2008]; often outcomes that cannot be billed are not coded. Similarly, outcomes may be up- or down-coded based on reimbursement preferences.

Automated extraction of patient outcomes from the rich EHR data source can improve quality of care. A recent article [Gandhi *et al.*, 2011] in the New England Journal of Medicine emphasizes the importance of completed patient problem lists, its role in avoiding medical mistakes and a way of deriving comprehensive lists using automated extraction algorithms. Automated outcome extraction can also serve as infrastructure for clinical trial recruitment, research, bio-surveillance and billing informatics modules.

Previous works have focused on using state of the art natural language processing (NLP) techniques for extracting patient outcomes from discharge summaries [Solt *et al.*, 2009; Pakhomov *et al.*, 2005; Friedman *et al.*, 2004]. Two common pipelines have typically been used. The first is a rule-based classification system such as that in Solt et al. [2009]. We build on their system and describe it in more detail subsequently. A second commonly used approach [Friedman *et al.*, 2004] is one that is less domain specific. Low-level modules such as a sentence boundary detector, tokenizer and spelling corrector are used to de-noise the clinical text. Modules trained on a clinical corpus are typically used. Next, higher-level natural language processing modules such as a parser, named entity recognizer and a part-of-speech analyzer are used to extract semantic structure from the sentences and recognize complications as salient phrases in the narratives.

Although these systems perform reasonably well, performance is limited by complex language structure in the dictated sentences [Meystre *et al.*, 2008]. First, clinical texts are ungrammatical and composed of short, telegraphic phrases. See examples in the discharge summary shown. Second, clinical narratives are rife with shorthand (abbreviations, acronyms, and local dialectal shorthand phrases). These shorthand lexical units are often overloaded (i.e., the same set of letters has multiple renderings); Liu et al. estimate that acronyms are overloaded about 33% of the time and are often highly ambiguous even in context [2001]. Third, misspellings abound in clinical texts, especially in notes without rich-text or spelling support. Finally, the presence of special characters and noise introduced due to transcription make word tokenization difficult. These problems occur in addition to the common problems in extracting semantics from complex natural language sentences.

Example discharge summary[1]

ADMISSION DIAGNOSES:<cr>1. A 31-6/7-week male infant, twin B.<cr>2. Prematurity.<cr>3. Possible sepsis.<cr><cr>DISCHARGE DIAGNOSES:<cr>1. A day of life number 34, ex-31-6/7-week male infant, now 36-5/7-weeks<cr>postconceptual age.<cr>2. Mild indirect hyperbilirubinemia. <cr>3. Resolved mild apnea of prematurity.<cr>4. Mild physiologicanemia. <cr>5. Immature retinae.<cr><cr>IDENTIFICATION: XXXX XXXX XXXX is a day of life number 34, ex-31-6/7-week<cr>male infant who was admitted secondary to prematurity and possible sepsis. His<cr>hospital course has been fairly unremarkable. He was advanced to full feeds. He<cr>has tolerated this well and has bouts of intermittent mild, indirect<cr>hyperbilirubinemia. He is now being discharged home with follow-up with Dr. XXX<cr>XXXX at the Palo Alto Medical Foundation.<cr><cr>BIRTH HISTORY: XXXX XXXX XXXX is a 1,640 gram product of a<cr>monochorionic-diamniotic concordant twin gestation pregnancy. He was born at<cr>Lucile Salter Packard Children's Hospital at 31-6/7-week gestation to a<cr>30-year-old, gravida 2, para 1-0-0-1 mother who received good prenatal care. <cr>Prenatal laboratories were as follows: Blood type B positive, antibody screen<cr>negative, hepatitis B surface antigen negative, rubella immune, RPR negative,<cr>gonorrhea and Chlamydia negative, HIV negative, and group B strep negative on<cr>XX/XX/XX. The mother had an initial Glucola screen that was increased at one<cr>hour, but the three-hour Glucola test was normal.<cr><cr>The pregnancy was otherwise uncomplicated until XX/XX/XX, when the mother awoke<cr>in a large pool of fluid at about 7:00 in the morning. This was rupture of<cr>membranes for twin A. She was not experiencing contractions or vaginal bleeding at<cr>that time.

Fetal movement was still notable. She was not having any dysuria or<cr>unusual vaginal discharge. She presented to the hospital that day and was admitted<cr>and treated for premature rupture of membranes. She was started on ampicillin and<cr>erythema on XX/XX/XX. A urine culture was checked and was negative. She was<cr>also given two doses of betamethasone on XX/XX/XX8, and XX/XX/XX. Pediatrics was<cr>called to the delivery of these twin infants. Twin A was delivered first in vertex<cr>vaginal delivery. Twin B was delivered second with rupture of membranes occurring<cr>only a few

minutes prior to delivery. He was delivered via breech extraction at<cr>1828 hours on XX/XX/XX. He was initially floppy and without cry or respiratory<cr>effort.

He was handed to the pediatrics team and was taken to the radiant warmer.<cr>He was warmed, dried, stimulated and suctioned, but remained floppy and without<cr>respiratory effort. His heart rate, however, was adequate. He was given masked<cr>CPAP for a couple of seconds and was then found to cry. The CPAP was removed and<cr>he was continued with stimulation.

His cry and respiratory effort continued to<cr>improve, as did his tone and grimace. Apgar scores were 3 at one minute and 9 at<cr>five minutes. He was then admitted to the Neonatal Intensive Care Unit for<cr>prematurity and suspected sepsis.<cr><cr>HOSPITAL COURSE BY SYSTEM:<cr>1. FLUIDS, ELECTROLYTES AND NUTRITION: The birth weight is 1,640 grams with a<cr>current weight of 2,561 grams. The infant was initially n.p.o. with IV fluids at<cr>80 mL/kg per day.

Total parenteral nutrition was started on day of life number<cr>one, which was discontinued on day of life number six with advancement to full<cr>feeds. His fluids were slowly increased to a total of 170 mL/kg per day by the<cr>first week of life. Maternal breast milk was initiated on day of life number two<cr>with a slow advance as available. He was advanced to full feeds by day of life<cr>number seven. He was changed to half strength human milk fortifier on XX/XX/XX,<cr>changed to ad lib feeds on XX/XX/XX, and changed to Enfacare 22 calories per<cr>ounce with maternal breast milk on XX/XX/XX. He is currently on Enfacare 22<cr>calories per ounce with maternal breast

milk nippling 75 to 100 mL every three to<cr>four hours. He is voiding and stooling appropriately.<cr><cr>His alkaline phosphatase peaked at 481 on XX/XX/XX. His last alkaline<cr>phosphatase level was 285 on XX/XX/XX. He has had normal electrolytes throughout<cr>his hospitalization.<cr><cr>2. CARDIOVASCULAR AND RESPIRATORY: The infant has been on room air since<cr>admission. He did have evidence of mild apnea of prematurity. However, he never<cr>required caffeine. He did originally have occasional, intermittent bradycardiac<cr>events with sleep, which were mild. His last bradycardia with sleep was on<cr>XX/XX/XX, for which he required very gentle

stimulation and was only a few<cr>seconds. The last time he had bradycardia with feeding was on XX/XX/XX, which<cr>self-resolved. At this point, he has had no further episodes and is safe for<cr>discharge.<cr><cr>3. HEMATOLOGIC: His admission hematocrit is 44.9\% with a recent hematocrit of<cr>32.7\%. He has not required any blood transfusions throughout his hospitalization.<cr><cr> He does have evidence of indirect hyperbilirubinemia with an initial peak bilirubin<cr>level early on of 9.7 on XX/XX/XX. He has been on and off phototherapy<cr>throughout his life with the phototherapy last being discontinued on XX/XX/XX.<cr>He has had a slow, rebound hyperbilirubinemia since that time with a bilirubin<cr>level of 9.6 on XX/XX/XX, and a bilirubin level of 10.6 on XX/XX/XX. His liver function tests have remained normal. He did have a G6PD<cr>sent on XX/XX/XX, which was normal, showing no evidence of deficiency.<cr><cr>4. INFECTIOUS DISEASE: He was started on antibiotics immediately after delivery<cr>of ampicillin and gentamicin for concern for infection. He had normal C-reactive<cr>proteins and his blood and cerebral spinal fluid cultures had remained negative.<cr>Therefore, his antibiotics were discontinued after 48 hours. He has since had no<cr>signs or symptoms of infection.<cr><cr>5. NEUROLOGICAL: A head ultrasound was done on XX/XX/XX, which was normal. He<cr>passed his ALGO hearing screen on XX/XX/XX.<cr><cr>6. OPHTHALMOLOGY: He had a retinopathy of prematurity examination number one<cr>done, secondary to pale red reflexes. This was done on XX/XX/XX, which showed<cr>immaturity of his retinae, zone III and a follow-up indicated at about ten days<cr>with Dr. XXX.<cr><cr>7. LABORATORY DATA:<cr>1. Hematocrit 32.7\% on XX/XX/XX.<cr>2. Total bilirubin level of 10.5 on XX/XX/XX.<cr>3. Newborn screen drawn on XX/XX/XX, was normal.<cr><cr>DIET: Ad lib maternal breast milk with Enfacare 22 calorie per ounce formula.<cr><cr>MEDICATIONS:<cr>1. Poly-Vi-Sol 1 mL p.o. daily.<cr>2. Fer-In-Sol 0.4 mL p.o. daily.<cr><cr>IMMUNIZATIONS:<cr>1. Hepatitis B vaccine given on XX/XX/XX.<cr>2. Synagis number one give on XX/XX/XX.<cr><cr>FOLLOW-UP CARE:<cr>1. Dr. XXX on Tuesday, XX/XX/XX.<cr>2. Follow-up with Dr. XXX for retinopathy of prematurity in ten days.<cr><cr>SPECIAL TESTING:<cr>1. Passed ALGO hearing screen on XX/XX/XX.<cr>2. Will need repeat eye examination in ten days.<cr><cr><cr> <cr><cr>D: XX/XX/XX 10:43 A<cr>T: XX/XX/XX 3:10 P<cr>I: XX/XX/XX 3:50 P<cr> <cr>DOC \#: 286761 DICT JOB \#: 001290294<cr><cr>

While a majority of the current work is focusing on building increasingly sophisticated language models, we take a complementary approach to this problem by incorporating simple cues extracted from structured EHR data when available. For example, treatments and medications are prescribed by clinicians specifically to manage patient complications; thus, presence or absence of relevant treatments can provide independent indicators to disambiguate cases where current NLP approaches fail. Similarly, clinical events such as a test being ordered or use of an equipment as a measurement device (e.g., ventilator) can also provide markers for specific complications. Thus, our proposed system combines a state-of-the-art extraction system with structured clinical event cues.

2.2 Automated Outcome Extraction

The task of automated outcome extraction entails identifying any complications that occur during the episode of care. In our case, our goal is to identify, for each infant, any complications that occurred during their entire length of time in the hospital.

For this purpose, we constructed a data set to evaluate our system. Two expert neonatologists formulated a list of all major complications observed in the NICU (Table 2.1). The data was annotated for these and any additional unlisted complications and subsequently reviewed by a team of three nurses and a physician. Overall, records of 275 premature infants born or transferred within the first week of life to the Stanford Lucile Packard Children Hospital's Neonatal Intensive Care Unit (NICU) after March 2008 and discharged before October 2009 were reviewed. We extracted discharge summaries, as well as laboratory reports of urine (188 reports) and blood cultures (590), radiology reports of ECHO (387) and head ultrasounds (534), medication events, and clinical events such as ventilator settings and tube placements. There were 628 unique complication-patient pairs marked as positive and 4872 complication-patient pairs marked as negative.

2.3 Methods

Recent work has shown the success of rule-based models in the domain of information extraction from clinical texts, in particular those employing hand-crafted string matching patterns to identify relevant lexical items and shallow semantic features [Solt *et al.*, 2009; Goldstein *et al.*, 2007]. While these models are not optimal on account of their inability

Complication	M	E	C	R
Respiratory Distress Syn (RDS)	X	X		
Sepsis	X			
Patent Ductus Arteriosus (PDA)	X			X
Bronchopulmonary Dyslapsia (BPD)	X	X		
Intraventricular Hemorrhage (IVH)				X
Died				
Pneumothorax (PNE)				
Adrenal Insufficiency (ADR)	X			
Coagnegative Stahylococcus (BCS)	X		X	
Necrotizing Enterocolitis (NEC)	X	X		
Bacterimia (BAC)	X		X	
Arrhythmia (ARR)	X			
Hydrocephalus (HYD)				X
Pulmonary Hemorrhage (PUL)				
Urinary Tract Infection (UTI)			X	
Adrenal Renal Failure (ARF)		X		
Pneumonia (PNA)				
Pulmonary Hypertension (PPHN)				
Seizure (SEI)	X			
Chronic Renal Failure (CRF)		X		

Table 2.1: List of complication-specific clinical features used. Complications are listed in order of decreasing frequency in our data set. Features are extracted from medications (M), clinical events (E), culture reports (C) and radiology reports (R). Overall, 33 clinical features are extracted.

to generalize, they usually have better performance than models which use general NLP strategies [Uzuner, 2009].

In constructing the language model component of our system, we built it based on the context-aware approach employed by the i2b2 Obesity Challenge winners, Solt et al. [2009]. To accurately evaulate the incremental contribution of incorporating structured information from the EHR, we replicate their model in our domain and use it as our language model baseline. Their approach aims to identify and categorize typical linguistic contexts in which patient disease outcomes are mentioned. The types of contexts which suggest a positive, negative, or uncertain result are fairly consistent within the domain of medical records, making it possible to engineer regular expressions that capture and categorize a majority of these mentions correctly. We describe below in detail the four basic types of language

based features that comprise our baseline language model.

2.3.1 Language Features

Disease Mentions: In addition to complication / disease names, this category includes patterns to capture abbreviations (e.g., *UTI* and *NEC*), alternate spellings (e.g., *haemorrhage* and *hemorrhage*), complication subclasses (e.g., *germinal matrix hemorrhage* and *intracranial hemorrhage* for IVH), and synonyms (e.g., *cardiac arrest* for arrhythmia.) Expert opinion was sought in increasing feature coverage. Querying the Unified Medical Language System (UMLS), a comprehensive ontology of biomedical concepts would be another way to derive terms that can improve cocerage. The deterministic model using just this set of rules maps most closely to the baseline binary classifier in Solt et al. [2009].

Negations: We use a Negex inspired strategy to identify both sentential and noun-phrase negations that indicate a negative result pertaining to one of the above disease name mentions. General patterns such as *no|never MENTION* and *(no|without) evidence of MENTION* are used across all disease types, but disease specific negation patterns are also allowed where appropriate, e.g., *r/o SEPSIS* (rule out sepsis).

Uncertainty modifiers: Uncertain contexts are identified by patterns of similar construction to the negation patterns but include templates such as *(possible|suspected) MENTION* and *history of MENTION*. It is important for the system to identify regions of uncertainty in order to avoid overvaluing many disease name mentions. Disease specific uncertainty patterns may also be used to recognize information that is most likely unrelated to patient outcome, e.g., *family death* or *pregnancy related UTI*.

Correlated Words and phrases: This final category of language features came from reviewing with experts words that showed high correlation with the outcome label. Similar to the process of automatically extracting symptoms, medications, and related procedures from the description of ICD-9 codes, we reviewed our data with medical professionals and arrived at pattern matches for names and abbreviations of relevant antibiotics, treatments (*antibiotics discontinued* for sepsis ruled out), symptoms (*PAC* for arrhythmia) and tests (*head ultrasound*).

A total of 285 language features were extracted. We experimented with several ways of combining these language features in our baseline model; we delay this discussion to the results section.

2.3.2 Clinical features

Structured information in the patient EHR can be extracted from sources other than the discharge summary, including records from diagnostic tests, medication and treatments administered. We refer to such features as *clinical* features. These features were developed with guidance from a neonatologist in two half hour sessions. For each complication, we listed various treatment options, medications provided, diagnostic tests used or other clinical events that are synonymous with the complication. Table 2.1 lists the various classes of clinical features that were used for each complication. Our overarching principle in implementing clinical features was simplicity of extraction. While more fine-tuned models can be built to improve sensitivity/specificity of features extracted from these different sources, our experiments show that even these relatively simple features are enough to significantly improve performance of the overall system. Our system includes a total of 33 clinical features.

Medications (M): The EHR stores the medication name, dosage, along with the time at which the medication was administered as structured events. Rules of the form *(medication name(s), minimum length of prescription)* were obtained from the neonatologist for all relevant complications. Such a rule is activated if a medication in the rule is administered to the infant for at least the minimum time. For example, for sepsis we have *(vancomycin & cefotaxime, 4 days)* as one of the rules. Each of these rules is modeled as a binary feature.

Clinical Events (E): For various clinical events associated with complications, we obtained rules of the form *(event name, minimum event duration, threshold event value)*. Events include therapies (for example, infants with respiratory distress syndrome are often on oxygen therapy represented as *(oxygen therapy, 1 day, N/A)*) as well as lab measurement (for example, extended increase in creatinine measurements is indicative of a renal malfunction in infants represented as *(creatinine, 2 days, 1.5)*). Each of these rules is modeled as a binary feature.

Culture Reports (C): Culture status is relevant to various complications. A vast majority of the cultures have a section that summarizes the result of the culture, where "No growth" is mentioned unless any bacterial growth is observed. We note that the presence of growth may be a result of a contaminant, which is further discussed in the unstructured text section of the report. For our current study, we do not make this correction. The result of each report is encoded as a binary response. The count over all reports for any given patient is modeled as a multinomial feature.

Radiology Reports (R): Our approach is based on prior work that placed second in a recent CMC challenge [Goldstein *et al.*, 2007]. For each type of report, we extract sections in decreasing order of relevance until a non-empty section is available. The section is parsed for indications of the complication or symptom mentioned in a positive, negated or uncertain context using the language rules described earlier. The counts for each type of report over all reports for any given patient are included in the model as a feature.

2.3.3 Learning Technique

For outcome label prediction, we use a penalized logistic regression model that combines all features. While a broad set of classifiers can be deployed, penalized logistic regression is known to perform well in the low data regime [Zhu and Hastie, 2004]. The weights for this model are learned using maximum likelihood regularized with ridge regression, which trades off fit to data with model complexity, as measured by the sum of the learned weights. That is, we optimize the training objective:

$$\arg\max_{\vec{w}} \sum_{d=1}^{D} \sum_{i=1:N} [-y_i^d \vec{w}^T(\vec{f_i}\vec{s}^d) + \ln(1+\exp(\vec{w}^T(\vec{f_i}\vec{s}^d)))] + \frac{1}{2\sigma^2}||\vec{w}||^2$$

where N is the number of training examples and d indexes each of the complications. $\vec{f_i}$ are the feature counts, $\vec{s}^d$ selects the features relevant to each disease. So, $\vec{s}_j^d = 0$ if the feature is extracted as being relevant to disease d and 1 otherwise. $y_i \in \{0, 1\}$ is the label of the ith example, $\vec{w}$ is the weight vector, and σ controls the magnitude of the ridge penalty.

Similar to [Crammer *et al.*, 2007], we develop *transfer* features that represent patterns that repeat across multiple complications and allow us to generalize from one label to another without having seen mentions of that feature in the training data. For example, *without sepsis* and *without pneumonia* both suggest the mention of the disease in a negated context. With a transfer feature *without (disease name)*, a negative weight learned from sepsis is applied in the context of pneumonia. Other examples of transfer features include *(disease name) ruled out*, *concern for (disease name)*. Of particular interest is the feature *PosMention (infrequent disease name)* which encodes sharing only amongst infrequently occurring complications. Complications like sepsis that are rampant in the population are discussed in almost every discharge summary and are ruled out using tests. Infrequent

complications are only discussed when the patients show complication-specific symptoms and thus, their mention alone is strongly correlated with having the complication. Each feature is encoded by a set of regular expressions that capture varying mentions in the data. The transfer features we used are listed in table 2.2. Weight sharing was similarly introduced for clinical features that were common to multiple complications (e.g., a positive blood culture is a diagnostic test used for both BAC and BCS).

Weight sharing can be implemented easily. We modify our learning objective as follows:

$$\arg\max_{\vec{w}} \sum_{d=1}^{D} \sum_{i=1:N} [-y_i^d \vec{w}^T(\vec{h}_i \vec{s}^d) + \ln(1 + \exp(\vec{w}^T(\vec{h}_i \vec{s}^d)))] + \frac{1}{2\sigma^2}||\vec{w}_l||^2 + \frac{1}{2\sigma^2}||\vec{w}_g||^2$$

where $\vec{w} = [w_l, w_g]$ and $\vec{h_i} = [f_i; f_i]$. The new feature vector $\vec{h_i}$ is formed by concatenating the matched features twice. $\vec{s}^d$ selects indices in $\vec{h}_i$ for features relevant for the disease. For example, "rule out sepsis" is a feature relevant to sepsis but not relevant to any other complication. Thus, the element corresponding to the "rule out sepsis" feature in $\vec{s}^d$ is 0 in all diseases except sepsis. w_l are complication-specific feature weights. w_g are weights for features that are shared between complications. Thus, the prediction for each data instance contains a contribution from the disease specific weights and the global weights. The inclusion of both transfer and disease specific features with a ridge penalty allows the model to learn specificity when there are large number of examples and generality for rare outcomes.

Concern for (disease name)
Possible | Suspected (disease name)
(disase name) ruled out
(disease name) resolved
Without | Not (disease name)
No | Never | None (disease name)
Negative for (disease name)
Significant (disease name)
History of (disease name)
Normal (disease name)
PosMention(infrequent disease name)

Table 2.2: Language transfer features.

2.4 Experiments and Results

We form a strong language model baseline by replicating for our domain, Solt et al. [2009], previous I2B2 challenge winners. We improve the baseline by learning weights for the rules in their model through weight sharing. We compare performance of the Integrated EHR model to the best baseline achieved using only language features.

2.4.1 Statistical Methods

We compute precision, recall, and F1 for each condition, and then compute overall precision, recall, and F1 using micro-averaging. Let tp, fp, tn, fn be the number of true positives, false positives, true negatives, and false negatives respectively. Then,

$$\text{Precision} := \frac{\text{tp}}{\text{tp} + \text{fp}} \qquad \text{Recall} := \frac{\text{tp}}{\text{tp} + \text{fn}} \qquad \text{F1} := 2\frac{\text{precision.recall}}{\text{precision} + \text{recall}}$$

All results reported are based on average test performance over 100 trials of randomized 70/30 train/test split. Significance values are computed using the bootstrap method on the 100 trials.

2.4.2 Baseline Language Model

Our aim in developing the language model (LM) was to maximize its performance, so as to best evaluate the incremental contribution obtained from the clinical features. Thus, the LM development was done on the entire dataset using random 70/30 train/test splits. The cross-validation parameter σ was set to 0.8 to optimize test performance of the LM in the hold-out set, and not subsequently adjusted for the inclusion of the clinical features.

We experimented with several approaches for combining the language features to derive a strong baseline (see Table 2.3). Similar to past winners [Goldstein *et al.*, 2007], we experimented with pre-fixed weighting schemes. A hand-tuned model was derived as follows: for a given patient-complication pair, all sentences from the discharge summary that matched language features for that complication were extracted. Each sentence was allowed at most one vote; a "Yes" vote was assigned if only disease mentions without negations or uncertainty matched the sentence or a "No" vote if any negated mentions of the disease matched. To combine all votes, a model that counted "No" votes twice as much as "Yes" votes gave the best results. *DLM*, deterministic language model, shows the performance of

this fixed weighting scheme model. *LLM*, learned language model, shows performance of the model with weights learned assuming the bag of all matched features using the learning technique described earlier. We also show contributions of component feature classes to the baseline by adding them incrementally. We use the LLM (all features), with F1 of 84.7, as the baseline for comparison with the EHR model.

Model	Feature Set	Prec.	Recall	F1
DLM	All features	73.5	86.1	79.4
LLM	Disease Mentions	88.7	72.8	79.9
	+ Negations	90.7	78.2	83.9
	+ Uncertain	90.8	77.8	83.7
	+ Correlated	90.6	79.5	**84.7**

Table 2.3: Baseline: language model performance.

2.4.3 Integrated EHR Model

The EHR model contains all language features as well as the clinical features. Unlike the language model, the clinical features did not have an iterative feature development phase and were determined apriori using expert medical knowledge. The model weights were trained using the standard bag of words assumption[2] with weight sharing for the transfer features as detailed earlier. In Table 2.4, we report test performance of the EHR model against our best language model. Overall, the EHR model with average F1 score of 88.3 performs significantly (p-value = 0.007) better than the language model. Additionally, the complications for which the EHR model does not outperform are those for which there were no clinical features included. From Table 2.1, note that for each complication, clinical features were extracted from only one or two sources.

2.4.4 Error Analysis

A post-hoc analysis of the results was done to understand the performance of our augmented model. We identify three distinct sources of error: (1) medical ambiguities, (2) feature error, i.e., failure of a language or clinical feature match on a specific instance, and (3) data

[2]All features are considered to be independent and the order in which the features appear are not taken into account. This assumption is commonly made in document processing.

	Language Model			EHR Model		
Comp	Pr.	Re.	F1	Pr.	Re.	F1
Respiratory Distress Syndome (RDS)	96.2	93.8	95.0	96.8	94.5	95.6
Sepsis (SEPSIS)	82.3	69.8	75.5	92.5	79.5	85.5
Patent Ductus Arteriosus (PDA)	92.4	85.6	88.9	94.7	87.0	90.7
Bronchopulmonary Dysplasia (BPD)	90.5	73.3	81.0	92.9	82.2	87.2
Intraventricular Haemorrhage (IVH)	92.9	79.0	85.4	96.2	78.5	86.5
DIED	95.0	93.9	94.5	94.7	93.7	94.2
Pneumothorax (PNE)	100.0	85.9	92.4	100.0	84.1	91.4
Adrenal Insufficiency (ADR)	90.4	56.8	69.8	91.4	64.2	75.4
Coagnegative Staphylococcus (BCS)	93.6	88.6	91.0	99.7	87.5	93.2
Necrotizing Enterocolitis (NEC)	76.5	59.5	66.9	74.6	61.5	67.4
Bacterimia (BAC)	69.6	11.3	19.5	100.0	68.6	81.3
Arrhythmia (ARR)	98.5	50.2	66.5	98.1	61.0	75.2
Hydrocephalus (HYD)	88.3	79.7	83.8	88.8	91.2	90.0
Pulmonary Hemorrhage (PUL)	100.0	99.5	99.8	100.0	90.5	95.0
Urinary Tract Infection (UTI)	59.0	58.5	58.7	55.7	57.0	56.3
Acute Renal Failure (ARF)	67.7	28.2	39.8	71.2	33.3	45.4
Pneumonia (PNA)	100.0	2.0	4.0	100.0	2.7	5.3
Pulmonary Hypertension (PPHN)	58.3	59.6	58.9	58.6	60.3	59.4
Seizure (SEI)	54.8	43.8	48.6	60.9	48.6	54.1
ALL	90.6	79.5	84.7	93.5	83.6	88.3

Table 2.4: Performance comparison between the language model and the EHR model. For visual clarity, the winning model, chosen based on F1, is underlined for each complication. For the outcome labels of death, penumonia, pneumothorax, pulmonary hemorrhage, and pulmonary hypertension, no clinical features were available.

extraction.

A significant source of error within the dataset is inherent ambiguity in the process of medical diagnosis. Beyond cases that are simply complex to code, there are patients for which even medical experts disagree about the underlying diagnosis. In example 1 below, we show sentences in the discharge summary related to sepsis. The infant was treated for sepsis and given a 7-day antibiotic course (corresponding text segment in the example has been emphasized by us for clarity) yet no positive blood culture was present. The annotators disagreed on the status of sepsis for this infant. Similarly, in example 2, the infant received respiratory support through CPAP, albeit for not too long. Annotators disagreed on the status of respiratory distress syndrome for the infant. The highest achievable F1 score in

our data with these examples included as errors is 96.3.

Example 1. Disagreement about status of Sepsis

> Presumed sepsis.6
> The infant was started on TPN and Intralipid on day of life number 1
> He also had some initial hypokalemia; however, this resolved with potassium infusion and an increase of potassium in his TPN
> INFECTIOUS DISEASES: He had an initial *7-day course of antibiotics* for his delivery

Example 2. Disagreement about status of RDS

> Apgars were 7 at 1 minutes (-1 tone, -1 respiratory, -1 color) and 8 at 5 minutes (-1 tone, -1 respiratory)
> CARDIOVASCULAR/RESPIRATORY: Initially, the patient required CPAP at 5;however, upon arrival to the NICU was rapidly weaned to room air

Feature errors in the language model (LM) can arise when context patterns fail to match because a lexical cue is separated from the disease mention by too much intervening text, but this turned out to be a relatively rare occurrence in our dataset. There were just four instances of error where syntactic parsing could have identified a modifier that was missed by regular expressions. A second type of language error, which occurs mainly with our most frequent complications, SEPSIS and RDS, are spans that contain atypical contexts and/or require inference. In the sentence, *"The workup was entirely negative and antibiotics were discontinued in approximately four days"*, there is no explicit mention of the complication, yet we can infer the patient most likely underwent a workup for sepsis. The addition of our 'Correlated Words' rule set helps mitigate these errors. In this case, for example, the rule *antibiotics discontinued after X hrs/days* correctly matched. In the full model, there were five errors of this type for RDS, one for SEPSIS, and one for PDA. The final type of feature error in the LM model is the most common, with at least ten instances in our complete dataset. It results when multiple mentions of a disease occur in conflicting contexts throughout the document or even within a single sentence. Temporal event resolution might improve performance in such cases.

Feature errors can also arise in clinical features, although less frequently due to the simplicity of their extraction. Such errors do occur mainly because combinations not covered by our feature set were administered. For example, cefotaxime or vancomycin are administered for at least four days when a patient has sepsis. However, some patients were switched from one to the other midway through their course, a feature not covered by our initial set.

A final source of error was due to errors in the data extraction software or due to incomplete records. For more than 10 patients, subsets of their clinical records such as ultrasound reports, culture reports or clinical events were missing in our extracted dataset. In example 3 below, the medication regime for the infant was non-traditional (midway through his prescription, his medications were changed). The only evidence for a UTI was the presence of a positive urine culture which was missing from his records.

Example 3. Missing data about urine culture

> Status post rule out postnatal sepsis.4
> Due to complications of the pregnancy, the momwas treated with Lovenox, low-dose aspirin, and vancomycin for positive group Bstrep status due to penicillin allergy
> He was therefore transferred to the NeonatalIntensive Care Unit once again to rule out sepsis
> He was initially treated onvancomycin and cefotaxime, however, when cultures of his *urine came back positivefor E coli*, this was switched to cefotaxime
> The patient's initial rule out sepsis following birth was negative
> Intravenous antibiotics werediscontinued on 11/14, and the patient started on prophylactic Keflex

Furthermore, for textual reports, occasionally missing word boundaries resulted in feature match errors. Overall, an improved clinical feature set with more coverage and better extraction software should bring performance much closer to the achievable F1-ceiling.

2.4.5 Improving human annotation

For experiments in the subsequent chapters of this thesis, we required gold standard labels for each patient in our included patient cohort. We found that using our system, we were able to significantly reduce human error and speed annotation by a human. We

compared annotations obtained from our system with annotations from the nurses. Our system also extracted evidence it was using to arrive at each annotation. When there were inconsistencies in the annotation, we presented both the records along with our label and the evidence contributing to that label. We found that our system was able to correct labels on approximately half of the records presented in the first iteration and a quarter in the next iteration. In example 3 above, the presence of a positive urine culture was overlooked by the annotator and accounted for once our model highlighted that evidence. Many of the remaining inconsistencies were due to ambiguities inherent in deducing the patient outcome in which case we accepted the majority annotation from the human annotators as the gold-standard. Additionally, we found that human errors were often a result of ignoring information contained in the EHR. As can be seen in the example discharge summary, these summaries are cumbersome to read. For many patients in the NICU, once admission notes, progress notes, discharge summary and other records are aggregated, their record can extend to well over 50 pages and therefore, it is easy to overlook critical information pertaining to a disease.

2.5 Discussion and Conclusion

We presented a system that rigorously validates an intuitive idea: integrating easy-to-extract structured information such as medications, treatments and laboratory results into current NLP-based information extraction systems can significantly boost coding accuracy. With the recent ubiquity of EHR systems, this data is broadly available in many contexts [Li *et al.*, 2008]. We believe this study opens several exciting avenues for future work.

There are several limitations to the clinical features used in our work. We make the assumption that all features contribute independently. When available, richer features that encode dependencies between multiple features can also help improve precision. For example, vancomycin and cefotaxime are given for all infection related complications, including bacterimia (BAC) and necrotizing enterocolitis (NEC). This results in positive feature contribution towards BAC even when the medication was administered for NEC. If the infant is on *NPO status* concurrent with medication administration, then likely the infant was given the medication for NEC and not BAC. Similarly, the medication hydrocortisone can be given for many reasons; however, if it is administered soon after a cortisol stimulation

test, then it is most likely given for adrenal insufficiency (ADR). Modeling such dependencies can improve feature specificities. Exploiting dependencies between the related tasks of predicting individual disease outcomes might improve performance; the application of conditional random fields (CRFs) [Lafferty *et al.*, 2001] towards this end would be an interesting extension to the current formulation.

Our current implementation is limited by the need to obtain expert opinion similar to other rule-based systems. While rule-based systems have been very successful in recent challenges [Uzuner, 2009], they are more cumbersome to scale due to the information acquisition bottleneck. Moreover, there may be valuable rules that did not occur to the expert in the development cycle. To remedy this, one can extract rules from known ontologies such as existing medication indication dictionaries [Burton *et al.*, 2008] or UMLS. Combined with techniques such as boosting [Friedman *et al.*, 1998], candidate rules can be constructed automatically. Such feature induction can also be integrated into an interactive system that uses experts to evaluate proposed rules for medical plausibility.

Our system mitigates shortcomings of current NLP techniques by encoding additional independent sources of information that provide reinforcement where entirely language based systems err. This has the additional benefit of building a more comprehensive case for each patient providing the health experts with a transparent system where the evidence supporting each decision can be verified holistically.

Chapter 3

Discovering Dynamic Signatures in Physiologic data

The task of discovering novel medical knowledge from complex, large scale and high-dimensional patient data, collected during care episodes, is central to innovation in medicine. This chapter, and the next addresses the task of discovery in the context of physiologic data from the bedside monitors.

In this chapter, we propose a method for exploratory data analysis and feature construction in continuous-valued physiologic time series. While our primary motivation comes from clinical data, our methods are applicable to other time series domain. Our method focuses on revealing shared patterns in corpora where individual time series differ greatly in their composition.

3.1 Introduction

Time series data is ubiquitous. The task of knowledge discovery from such data is important in many scientific disciplines including patient tracking, activity modeling, speech and ecology. For example, in our domain of seeking to understand disease pathogenesis from physiologic measurements (e.g., heart rate signal shown in figure 3.1), several interesting questions arise. Are there any repeating patterns or signatures in this data? How many such signatures exist and what their characteristics might be? Furthermore, are there collections of signatures that co-occur and are indicative of the underlying (disease) state? Such questions arise in other domains as well including surveillance and wild-life monitoring. In

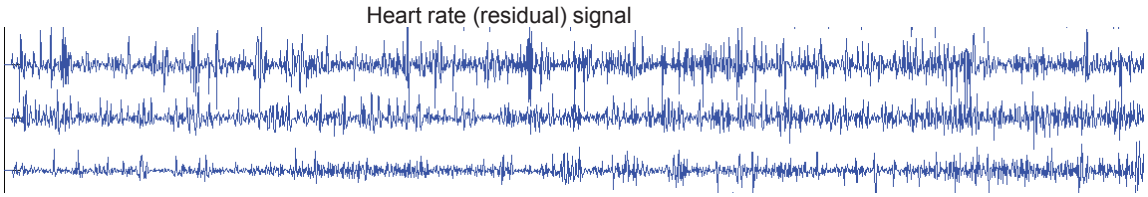


Figure 3.1: Heart signal (mean removed) from three infants in their first few hours of life.

such domains where the outcome of interest (e.g., health status) is difficult to measure directly and surrogate measurements are made instead (e.g., physiological variables), latent (hidden) variable models are a natural choice for knowledge discovery. Different diseases might be associated with multiple latent states that each generate data with distinct physiologic characteristics. Structure discovered with such a model can help reveal how diseases manifest, uncover novel disease associations, and highlight relationships between diseases.

In many temporal domains, individual series show significant variability and an a priori breakdown of data into distinct sets is unclear. In clinical data, for example, two patients are rarely alike; they may suffer from different sets of diseases and to varying extents. Traditional generative models for time series data, such as switching Kalman filters [Bar-Shalom and Fortmann, 1987] or mixtures of such models [Fine *et al.*, 1x998], assume the data to be generated from a discrete set of classes, each specifying the generation of a homogeneous population of i.i.d. time series. To see the shortcoming of such an approach, in our example, the patient state over time transitions over a large set of latent states (coughing, wheezing, sleeping and so on). Generation of all series from a single transition matrix over the set of latent states assumes that all series express these latent states in the same proportion (on expectation). But, in reality, different patients express these states in radically different proportions, depending on their combination of diseases and other physiological factors. While mixture models (inducing a distribution over different dynamic models) can generate additional variability, the set of possible combinations can grow combinatorially large. And, thus, a pre-imposed partition of the space of patients into a fixed number of classes limits our ability to model instance-specific variability.

Hierarchical Bayesian modeling [Kass and Steffey, 1989; Gelman *et al.*, 1995] has been proposed as a general framework for modeling variability between individual "units". As an example of this framework, in the domain of natural language processing, Latent Dirichlet

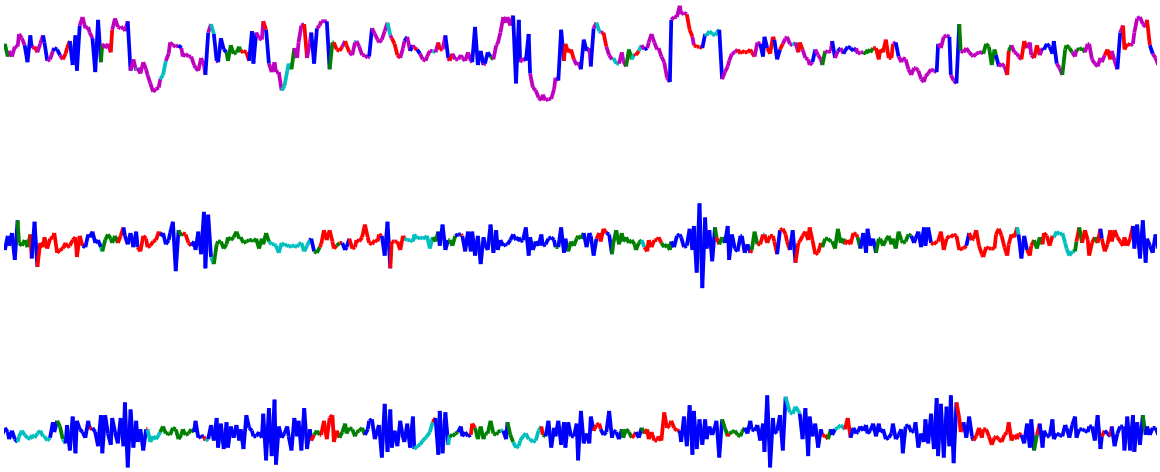


Figure 3.2: Example of three time series with shared signatures. Segments of each distinct color are generated by the same function, and thereby are instances of the same signature or *word*. Signatures here correspond to autoregressive functions. The choice of function used at any given time depends on the latent *topic* at that time. While the three series differ greatly in their composition, they contain shared structure to varying extents.

Allocation (LDA) [Blei *et al.*, 2003] has found success as a representation for uncovering the underlying structure of document corpora. Each document is associated with its own distribution over latent variables called topics, each of which is shared across the population and defines a distribution over words. Analogously, in our application[1], an individual patient maintains its own distribution over both latent (disease) topics and transitions between them. Each topic defines a distribution over temporal signatures (physiologic symptoms) observed in the time series and these behaviors play the role of words. However, unlike text data, in continuous-valued time series data, the notion of a word is non-obvious. A word could be specified as a segmented window of the data itself, but this allows for little compression, as most continuous-valued time series segments, unlike discrete text segments, do not repeat exactly. Our proposed model uses a more flexible representation of a word that specifies a parametric function to generate the temporal dynamics for the duration of that word. For example, in figure 3.2, autoregressive functions are used for generating the temporal dynamics. Each distinct color can be likened to a word and therefore, there are

[1]Our model is a more general instance of Hierarchical Bayes than LDA which models only discrete data. The analogy to LDA is made primarily to provide the readers a familiar overview of our model.

five "words" in this corpora. Moreover, the duration of the word also does not need to be fixed in advance, and as shown may vary from one occurrence to another. Hence, our model also postulates word boundaries.

In our approach, words are selected from an infinite dimensional latent space that corresponds to the possible real-valued instantiations to the parameters of the functions that generate the data. Naive sampling in this infinite-dimensional space given the data will result in no sharing of words across topics [Teh *et al.*, 2006]. For knowledge discovery tasks, sharing of words across topics is particularly desirable, as it allows us to uncover relationships between different latent states. For example, one can infer which diseases are physiologically similar based on the extent to which they share words. To enable sharing, we utilize *hierarchical Dirichlet processes* (HDPs) [Teh *et al.*, 2006], designed to allow sharing of mixture components within a multi-level hierarchy. Thus, our model discovers words and topics shared across the population while simultaneously modeling series-specific dynamics.

The chapter is structured as follows: we first give background on the existing building blocks of Dirichlet Processes and HDPs used in our model. We then describe the time series topic model (TSTM), a flexible hierarchical latent variable model for knowledge discovery in time series data, especially useful for domains when between series variability is significant. Next, we describe related work in models for processing continuous time-series data. Following this, we provide a block Gibbs sampler for TSTM. We present results on our target application of analyzing physiological time series data. We demonstrate usefulness of the model in constructing features within a supervised learning task. We also qualitatively evaluate the model output and derive new clinical insights that led to the development of a state-of-the-art personalized risk stratification score for morbidity in infants described in Chapter 5.

3.2 Background

Below, we briefly define Dirichlet Process and the Hierarchical Dirichlet Process. Though several texts have described these distributions in great detail before, for the sake of being comprehensive, we describe key properties here that give intuition about their use as priors on mixture models. The text for the following subsections is adapted from Fox et al. [2007].

3.2.1 Dirichlet Processes

The Dirichlet Process (DP) is commonly used as a prior on the parameters of a mixture model with a random number of components. A DP is a distribution on probability measures on a measurable space Θ. This stochastic process is uniquely defined by a base measure H on Θ and a concentration parameter γ. Consider a random probability measure $G_o \sim DP(\gamma, H)$. The DP is formally defined by the property that for any finite partition $\{A_1, \cdots, A_K\}$ of Θ,

$$(G_0(A_1), \cdots, G_0(A_K)) \mid \gamma, H \ \sim \ \mathrm{Dir}(\gamma H(A_1), \cdots, \gamma H(A_K))$$

That is, the measure of a random probability distribution $G_o \sim DP(\gamma, H)$ on every finite partition of Θ follows a finite-dimensional Dirichlet distribution [Ferguson, 1973]. A more constructive definition of the DP was given by Sethuraman [1994]. He shows that $G_0 \sim \mathrm{DP}(\gamma, H)$, a sample drawn from the DP prior, is a discrete distribution because, with probability one:

$$G_0 = \Sigma_{k=1}^{\infty} \beta_k \delta_{\theta_k} \tag{3.1}$$

where $\theta_k \sim H$. The sampling of β_k follows a *stick-breaking construction* defined as:

$$\beta_k = \beta'_k \prod_{l=1}^{k-1} (1 - \beta'_l) \qquad \beta'_k \sim \mathrm{Beta}(1, \gamma) \tag{3.2}$$

Essentially, we have divided a unit stick into lengths given by the weights β_k. β_k is a random proportion β'_k of the remaining stick after the previous $(k-1)$ weights have been defined. Generally, this construction is denoted by $\beta \sim \mathrm{GEM}(\gamma)$.

To give intuition for how the DP is used as a prior on the parameters of a mixture model with a random number of components, consider draws from H to be the description of candidate cluster centers. The weights β_k define the mixing proportions. γ controls the relative proportion of the mixing weights, and thus determines the model complexity in terms of the expected number of components with significant probability mass.

To see why the DP as a prior induces clustering, we visit another property of the DP, introduced by Blackwell and MacQueen [1973]. Consider a set of observations $\{y_i\}$ sampled from G_0. Consider z_i to be the variables that select for each data observation y_i the unique

value θ_k that the observation is sampled from i.e. $y_i \sim f(\theta_{z_i})$. Let K be the number of unique θ_k values that have data observations $y_1, \cdots, y_N$ associated with them.

$$p(z_{N+1} = z | z_1, \cdots, z_N, \gamma) = \frac{\gamma}{N+\gamma}\mathcal{I}(z = K+1) + \frac{1}{N+\gamma}\sum_{k=1}^{K}\sum_{i=1}^{N}\mathcal{I}(z_i = k)\mathcal{I}(z = k)$$

In other words, z_{N+1} samples its value based on how frequently these values have been used by previously sampled data observations and is more likely to sample a frequently sampled value. Thus, we see that the DP has a reinforcement property that leads to a clustering of the data. This property is essential in deriving finite and compact models.

Finally, we can also obtain the DP mixture model as the limit of a sequence of finite mixture models. It can be shown under mild conditions that if the data were generated by a finite mixture, then the DP posterior is guaranteed to converge (in distribution) to that finite set of mixture parameters [Ishwaran and Zarepour, 2002a]. Let us assume that there L components in a mixture model and we place a finite-dimensional Dirichlet prior on these mixture weights:

$$\beta | \gamma \sim \text{Dir}(\gamma/L, \cdots, \gamma/L) \tag{3.3}$$

Let $G_0^L = \sum_{k=1}^{L} \beta_k \delta_{\theta_k}$. Then, it can be shown [Ishwaran and Zarepour, 2000; 2002b] that for every measurable function f integrable with respect to the measure H, this finite distribution G_o^L converges weakly to a countably infinite distribution G_0 distribution according to a Dirichlet process. Similar to Fox et al., [2007], we use this truncation property in the development of our block Gibbs sampler.

3.2.2 Hierarchical Dirichlet Processes

There are many scenarios where groups of data are thought to be produced by related, yet distinct, generative processes. For example, in our target application of physiologic monitoring, different diseases and syndromes likely share physiologic traits, yet data for any single disease should be grouped and described by a similar but different model from that of another disease. Similarly, in document modeling, news articles may share topics in common. Yet, documents published in, say, the New York Times should be grouped and described by a similar but different model from that of the Wall Street Journal. Similarly, in surveillance, different activity trajectories may contain activities in common. Yet, data

collected at different times of the day should be modeled as different but similar groups.

The Hierarchical Dirichlet Process [Teh *et al.*, 2006] extends the DP to enable sharing in such scenarios by taking a hierarchical Bayesian approach: a global Dirichlet process prior $DP(\eta, G_0)$ is placed on Θ and group-specific distributions are drawn from a the global prior $G_j \sim DP(\eta, G_0)$, where the base measure G_o acts as an "average" distribution across all groups. When the base measure $G_0 \sim DP(\gamma, H)$ itself is distributed according to a Dirichlet process, the discrete atoms θ_k are shared both within and between groups. If the base measure G_0 were instead fixed and absolutely continuous with respect to Lebesgue measure, there would be zero probability of the group-specific distributions having overlapping support.

More formally, draws $G_d \sim \mathrm{DP}(\eta, G_o)$ from an HDP can be described as

$$G_0 = \Sigma_{k=1}^{\infty} \beta_k \delta_{\theta_k} \qquad \beta_k \sim \mathrm{GEM}(\gamma), \quad \theta_k \sim H \tag{3.4}$$

$$G_d = \Sigma_{k=1}^{\infty} \Sigma_{t=1}^{\infty} \hat{\beta}_{jt} \delta_{\theta_{jt}} I(\theta_{jt} = \theta_k) \qquad \hat{\beta}_{jt} \sim \mathrm{GEM}(\eta), \quad \theta_{jt} \sim G_0 \tag{3.5}$$

Essentially, since G_d samples its values θ_{jt} from a discrete distribution, any given atom θ_k from the base distribution can be sampled more than once. The corresponding weight for that atom θ_k in G_d is computed by aggregating the sampled weights $\hat{\beta}_{jt}$ for all atoms $\theta_{jt} = \theta_k$.

As with the DP, the HDP mixture model has an interpretation as the limit of a finite mixture model. Placing a finite Dirichlet prior on the global distribution induces a finite Dirichlet prior on the group-specific distribution and as $L \to \infty$, this model converges in distribution to the HDP mixture model [Teh *et al.*, 2006]:

$$\beta | \gamma \sim \mathrm{Dir}(\gamma/L, \cdots, \gamma/L) \tag{3.6}$$

$$\phi_j | \eta, \beta \sim \mathrm{Dir}(\eta\beta_1, \cdots, \eta\beta_L) \tag{3.7}$$

Our block Gibbs sampler for performing inference in the TSTM exploits this truncation property.

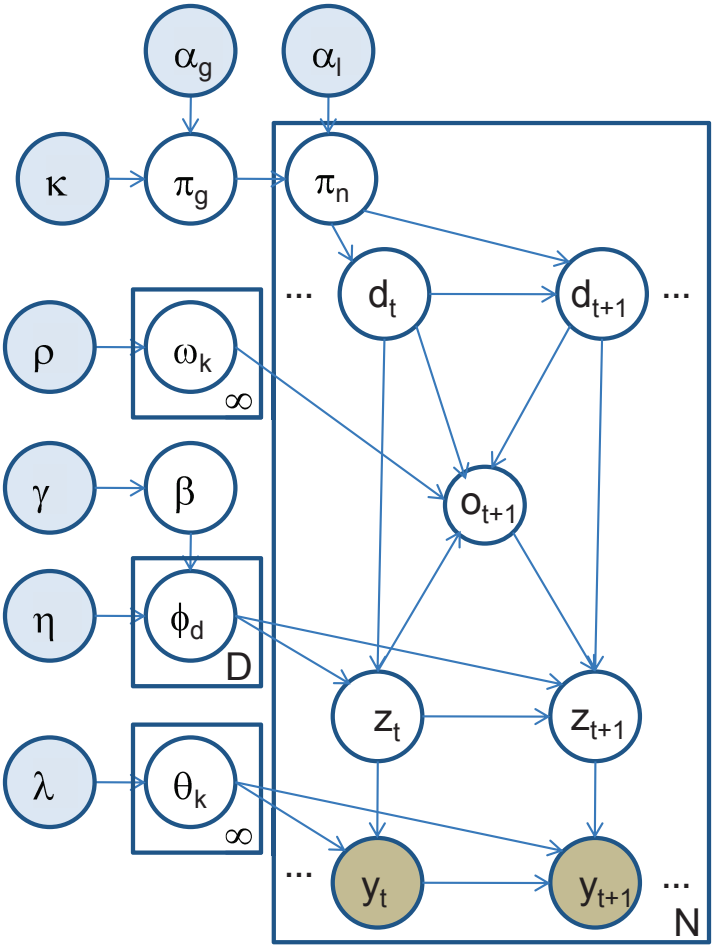


Figure 3.3: Graphical representation of the Time Series Topic Model (TSTM).

3.3 Time Series Topic Model

Time Series Topic Model (TSTM) is a 4-level hierarchical Bayesian model. It makes the assumption that there is an underlying fixed set of topics that is common to the heterogeneous collection of time series in the corpus. A topic is a distribution over the vocabulary of all words present in the corpus. An individual time series is generated by first choosing a series-specific transition matrix over the topics. To sample each "word": sample a topic, and then sample a word from that topics' distribution over words.

As discussed above, unlike discrete sequence data (e.g., text), in time series data the

Symbol	Description
D	Number of topics.
$\phi_{1:D}$	Topics 1 through D.
ϕ_d	The dth topic.
π_g	Global topic transition matrix.
π_n	Series specific topic transition matrix.
$f_k(.;\theta_k)$	parametric functions that generate the data; k indexes these functions. Also, refered to as the kth word.
y_t	Data observed at time t.
$\mathcal{I}(E)$	Indicator function where E is the event.
$1:D$	Abbreviation for 1 through D.

Table 3.1: Notation.

features to be extracted are often not structurally obvious (see figure 3.1). Pre-segmenting the sequence data into "words" does not offer sufficient flexibility to learn from the data, especially in the realm of exploration for knowledge discovery. Thus, TSTM discovers words from an infinite-dimensional parametric function space while simultaneously learning topics and series-specific evolution parameters.

We describe below each of the TSTM components. We begin by giving a brief overview of the random variables in the model and then describe the generation process (in a bottom-up fashion). We define notation that is commonly used in this chapter in table 3.1.

3.3.1 Overview of the model variables

Random variable y_t denotes the observation at a given time t (see figure 3.3 for the graphical model). z_t is a latent variable that indexes the "word" used to generate the data observed at that time. $d_t \in \{1, \cdots, D\}$ tracks the latent topic at any given time. Binary variables o_t control the word length. The word at a given time t, z_t is generated from the topic distribution ϕ_{d_t}. Each series has a series specific topic transition matrix π_n from which d_t is sampled at each time t. The matrices π_n are sampled from a global topic transition matrix π_g.

3.3.2 Generative Process

Data generation model: Let $\mathcal{F} = \{f(. : \theta) : \theta \in \Theta\}$ be the set of all possible data generating functions and $f_k = f(. : \theta_k)$ be the fuction indexed by k. We assume that the continuous-valued data y_t at time t is generated using a function f_k These functions take as inputs $\vec{x}_t$, values dependent on current and previous time slices, and generate the output as $y_t = f(\vec{x}_t; \theta_k)$, denoted as f_k. f_k, an expressive characterization of the time series dynamics, can be thought of as the kth word in the time-series corpus vocabulary. The parameterization of f_k depends on the choice of the observation model. Below, we describe the Vector Autoregressive Process (VAR) observation model. VARs have been used extensively for temporal modeling in numerous domains, including medical time series of fMRI, EEG and physiologic data [Williams *et al.*, 2005a]. We use this observation model for our target application and refer to functions discovered by applying TSTM to physiologic data as *dynamic signatures*.

Depending on the data, other observation models (such as the mixture model emissions utilized in [Fox *et al.*, 2007]) can be used instead within TSTM.

In an order p autoregressive process, given a function f_k with parameters $\{A^k, V^k\}$, observed data y_t is assumed to be generated as:

$$\vec{y}_t = A^k X_t^T + \vec{v}_t \qquad v_t \sim \mathcal{N}(0, V^k)$$

and $\vec{y}_t \in \mathcal{R}^m$ for an m-dimensional series. The inputs, $X_t = [\vec{y}_{t-1}, \ldots, \vec{y}_{t-p}]$. Parameters $A^k \in \mathcal{R}^{m \times p}$, and V^k is an $m \times m$ positive-semidefinite covariance matrix. The kth word then corresponds to a specific instantiation of the function parameters $\{A^k, V^k\}$. For TSTM, we want the words to persist for more than one time step. Thus, for each word, we have an additional parameter ω_k that specifies the mean length of the word as $1/\omega_k$. We describe how ω_k is used in data generation, below. For any $f_k \in \mathcal{F}$, we denote the function parameters more generally by $\vec{\theta}_k \in \Theta$.

Dynamics of words and topics: Given the words ($\mathcal{F}$), topics ($\phi_{1:D}$, D is the maximum number of topics) and series-specific transition matrices (π_n), the series generation is straightforward. For each time slice $t \in 1, \cdots, T$,

1. generate the current latent topic state given the topic at previous time-step, $d_t \sim \text{Mult}(\pi_n^{d_{t-1}})$,

2. generate the switching variables o_t, which determine whether a new word is selected. A new word is always generated ($o_t = 0$) if the latent state has changed from the previous time step; otherwise, o_t is selected from a Bernoulli distribution whose parameter determines the word length. Thus, $o_t \sim \mathcal{I}(d_t = d_{t-1})\text{Bernoulli}(\omega_{z_{t-1}})$, where I is the indicator function.

3. the identity of the word to be applied is generated; if $o_t = 1$, we have $z_t = z_{t-1}$, otherwise $z_t \sim \text{Mult}(\phi_{d_t})$.

4. the observation given the temporal function index z_t is generated as $y_t \sim f(x_t; \theta_{z_t})$.

The series specific topic transition distribution π_n is generated from the global topic transition distribution π_g. To generate π_n, each row i is generated from $\text{Dir}(\alpha_l \pi_g^i)$, where π_g^i is the ith row of the global topic transition distribution. Hyperparameter α_l controls the degree of sharing across series in our belief about the prevalence of latent topic states. A large α_l assigns a stronger prior and allows less variability across series. Given hyperparameters α_g and κ, $\pi_g^i \sim Dir(\alpha_g + \kappa\delta_i)$. κ controls the degree of self-transitions for the individual topics.

Word and Topic descriptions: To uncover the finite data generating parametric function set $\mathcal{F}$ where these functions are shared across latent topics ϕ_d, we use the hierarchical Dirichlet process (HDP) [Teh *et al.*, 2006]. Thus,

$$\phi_d \sim DP(\eta, \beta), \quad \beta \sim \text{GEM}(\gamma), \quad \theta_k \sim H \tag{3.8}$$

First, we define the base distribution H. Similar to [Fox *et al.*, 2009], we use a matrix-normal inverse-Wishart prior on the parameters $\{A^k, V^k\}$ of the autoregressive process and a symmetric Beta prior on ω_k as our base measures H. ϕ_d and β are easily generated using the truncation property, described in detail as part of the inference algorithm in section 3.5.

While we do not use the stick-breaking representation in our derivation of the inference algorithm, it is instructive to see how the HDP induces shared words between topics in the TSTM. Draws from H yield candidate words or data generating functions denoted by atoms δ_{θ_k} in Eq. 3.4. By associating each data sample y_t (time points in the series) through the latent variables z_t with a specific data generation function, the posterior distribution yields a probability distribution on different partitions of the data. The mixing proportion (the weights for each θ_k in Eq. 3.4) in the posterior distribution is obtained from aggregating

corresponding weights from the prior and the assigned data samples.

Since measures G_d in Eq. 3.5 are sampled from $\mathrm{DP}(\eta, G_0)$ with a discrete measure as its base measure, the resulting topic distributions ϕ_d have a non-zero probability of regenerating the same data generating functions θ_k, thereby sharing functions between related topics.

3.4 Related Work

An enormous body of work has been devoted to the task of modeling time series data. Probabilistic generative models, the category to which our work belongs, typically utilize a variant of a switching dynamical system [Bar-Shalom and Fortmann, 1987; Fine *et al.*, 1x998], where one or more discrete state variables determine the momentary system dynamics, which can be either linear or in a richer parametric class. The autoregressive hidden Markov model (AR-HMM) is one such example. Observations at any given time are generated via an autoregressive process. A hidden Markov model selects the choice of autoregressive process used at that time. However, these methods, as discussed above, typically utilize a single model (as is the case in an AR-HMM) for all the time series in the data set, or at most define a mixture over such models, using a limited set of classes. These methods are therefore unable to capture significant individual variations in the dynamics of the trajectories for different patients, required in our data.

Recent work by Fox and colleagues [Fox *et al.*, 2009; 2007; 2008] uses nonparametric Bayesian models for capturing the generation of continuous-valued time series. Works [Fox *et al.*, 2007; 2008] have utilized hierarchical Dirichlet process priors for inferring the number of features in hidden Markov models and switching linear dynamical systems but, akin to our discussion of [Bar-Shalom and Fortmann, 1987] above, these models do not explicitly represent variability across exemplar series. Conceptually, the present work is most closely related to BP-AR-HMMs [Fox *et al.*, 2009], which captures variability between series by sampling subsets of words (AR processes) specific to individual series. However, unlike TSTM, BP-AR-HMM does not have a mechanism for modeling the high-level topics that hierarchically capture structure in the collection of words. We show example results from the BP-AR-HMM in the results section to further elucidate the benefits of the generation mechanism of TSTM over BP-AR-HMM. Temporal extensions of LDA [Wang *et al.*, 2008; Wang and McCallum, 2006] model evolution of topic compositions over time in text data but not continuous-valued temporal data.

A very different approach to analyzing time series data is to attempt to extract features from the trajectory without necessarily constructing a generative model. For example, one standard procedure is to re-encode the time series using a Fourier or wavelet basis, and then look for large coefficients in this representation. However, the resulting coefficients do not capture signals that are meaningful in the space of the original signal, and are therefore hard to interpret. Features can also be constructed using alternative methods that produce more interpretable output, such as the work on sparse bases [Lee *et al.*, 2009]. However, this class of methods, as well as others [Mueen *et al.*, 2009], require that we first select a window length for identifying common features, whereas no such natural granularity exists in many applications. Moreover, none of these methods aims to discover higher level structure where words are associated with different "topics" to different extents.

3.5 Approximate Inference

Several approximate inference algorithms have been developed for mixture modeling using the HDP; see [Teh *et al.*, 2006; Fox *et al.*, 2007; Kurihara *et al.*, 2007] for a discussion and comparison. We use a block-Gibbs sampler that relies on the *degree L weak limit* approximation presented in [Ishwaran and Zarepour, 2002b]. This sampler has the advantage of being simple, computationally efficient and shows faster mixing than most alternate sampling schemes [Fox *et al.*, 2007]. The block-Gibbs sampler for TSTM proceeds by alternating between sampling of the state variables $\{d_t, z_t\}$, the model parameters, and the series specific transition matrices.

We detail the update steps of our block-Gibbs inference algorithm below. To briefly describe new notation used below, n indexes individual series. We drop the index n when explicit that the variable refers to a single series. We drop sub-indices when all instances of a variable are used (e.g., $z_{1:N,1:T_n}$ is written as z for short).

Sampling latent topic descriptions β, ϕ_d: The DP can also be viewed as the infinite limit of the order L mixture model [Ishwaran and Zarepour, 2002b; Teh *et al.*, 2006]:

$$\beta|\gamma \sim \mathrm{Dir}(\gamma/L, \cdots, \gamma/L)$$

$$\phi_d \sim \mathrm{Dir}(\eta\beta) \qquad \theta_k \sim H$$

We can approximate the limit by choosing L to be larger than the expected number of words in the data set. The prior distribution over each topic-specific word distribution is then:

$$\phi_d|\beta,\eta \sim \text{Dir}(\eta\beta_1,\cdots,\eta\beta_L)$$

Within an iteration of the sampler, let $m_{d,l}$ be the counts for the number of times $z_{n,t}$ sampled the lth word[2] for the dth disease topic; that is, let

$$m_{d,l} = \sum_{n=1:N}\sum_{t=1:T_n} \mathcal{I}(z_{n,t}=l)\mathcal{I}(d_{n,t}=d)\mathcal{I}(o_{n,t}=0)$$

and

$$m_{\cdot,l} = \sum_{d=1:D} m_{d,l}$$

The posterior distribution for the global and individual topic parameters is:

$$\beta|z,d,\gamma \sim \text{Dir}(\gamma/L + m_{\cdot,1},\cdots,\gamma/L + m_{\cdot,L})$$

$$\phi_{d'}|z,d,\eta,\beta \sim \text{Dir}(\eta\beta_1 + m_{d',1},\cdots,\eta\beta_L + m_{d',L})$$

Sampling word parameters ω_l and θ_l: Loosely, the mean word length of the lth word is $1/\omega_l$. A symmetric Beta prior with hyperparameter ρ, conjugate to the Bernoulli distribution, is used as a prior over word lengths. The sufficient statistics needed for the posterior distribution of ω_l are the counts:

$$\bar{c}_{l,i} = \sum_{n=1:N}\sum_{t=1:T_n} \mathcal{I}(d_{n,t}=d_{n,t-1})\mathcal{I}(z_{n,t-1}=l)\mathcal{I}(o_{n,t}=i)$$

where $i \in \{0,1\}$, representing the number of time steps, across all sequences, in which the topic remained the same, the word was initially l, and the word either changed ($o_{n,t}=1$) or not ($o_{n,t}=0$). Thus, $\omega_l|\bar{c}_{l,\cdot},\rho \sim \text{Beta}(\rho/2+\bar{c}_{l,1},\rho/2+\bar{c}_{l,0})$.

[2]Within this approximation, words are ordered such that all words that are observed in the corpus are assigned indices less than L. Thus, l indexes the lth observed word, which can correspond to different parameter instantiations over different iterations.

For sampling the AR generating function parameters, note that, conditioned on the mode assignments z, the observations $y_{1:T,1:N}$ can be partitioned into sets corresponding to each unique $l \in L$. This gives rise to L independent linear regression problems of the form $Y^l = A^l X^l + E^l$ where Y^l is the target variable, with observations generated from mode l, stacked column-wise. X^l is a matrix with the corresponding r lagged observations and E^l is the corresponding noise matrix. The parameters A^l and V^l are sampled from the posterior given conjugate priors of the Matrix-Normal Inverse-Wishart, similar to [Fox *et al.*, 2009].

Sampling global and series-specific transition matrices, π_g and π_n: Since the number of topic states D is known, and we use conjugate priors of Dirichlet distribution for each row of the transition matrix, the posterior update simply involves summing up counts from the prior and the data. The relevant count vectors are computed as $c^i_{n,k} = \sum_{t=1}^{T_n} \mathcal{I}(d_{n,t-1} = i)\mathcal{I}(d_{n,t} = k)$ and $c^i_k = \sum_{n=1}^{N} c^i_{n,k}$ which aggregates over each series. $\vec{c^i} = \{c^i_1, \cdots, c^i_D\}$ and i indexes a row of the transition matrix:

$$\pi^i_g | d, \alpha_g, \kappa \sim \text{Dir}(\alpha_g + \kappa\delta_i + \vec{c^i})$$

$$\pi^i_n | \pi_g, d, \alpha_l \sim \text{Dir}(\alpha_l \pi^i_g + c^i_{n,1:D})$$

Sampling state variables: If all model parameters (topic and word descriptions) are specified, then one can exploit the structure of the dependency graph to compute the posterior over the state variables using a single forward-backward pass. This is the key motivation behind using block Gibbs. The joint posterior can be computed recursively. Forward sampling is used to sample the variables in each time slice given the samples from the previous time slice as $P(z_{1:T}, d_{1:T} | y_{1:T}, \vec{\pi}) = \prod_t P(z_t, d_t | z_{t-1}, d_{t-1}, y_{1:T}, \vec{\pi})$. Top-down sampling is used within a given time slice.

Let $\vec{\pi}$ represent the vector of all model parameter values $\{\pi_{1:N}, \omega_{1:L}, \theta_{l:L}, \phi_{1:D}\}$ instantiated in the previous Gibbs iteration. Since state variables for individual time series n can be sampled independently from the posterior, we drop this index and represent $\vec{s}_t = \{d_t, z_t, o_t\}$ as the state variables in time slice t for any given series. When obvious, we drop mention of the relevant model parameter to the right of the conditioning bar.

The joint posterior is:

$$P(z_{1:T}, d_{1:T}|y_{1:T}, \vec{\pi}) = \prod_t P(z_t, d_t|z_{t-1}, d_{t-1}, y_{1:T}, \vec{\pi})$$

To use forward sampling to sample variables in each time slice given samples for the previous time slice, we compute $P(\vec{s}_t|\vec{s}_{t-1}, y_{1:T}, \vec{\pi})$. This can derived recursively as:

$$P(z_t, d_t|z_{t-1}, d_{t-1}, y_{1:T}, \vec{\pi}) = \frac{P(z_t, d_t|z_{t-1}, d_{t-1}, \vec{\pi}) g(y_{t:T}|z_t, d_t, \vec{\pi})}{\sum_{z_t, d_t} P(z_t, d_t|z_{t-1}, d_{t-1}, \vec{\pi}) g(y_{t:T}|z_t, d_t, \vec{\pi})} \tag{3.9}$$

The first term in the numerator is:

$$\begin{aligned} P(z_t, d_t|z_{t-1}, d_{t-1}, \vec{\pi}) &= P(d_t|d_{t-1}, \pi_n^{d_{t-1}})(P(o_t = 1|\omega_{z_{t-1}}, z_{t-1})\mathcal{I}(z_{t-1} = z_t) + \\ &\quad P(o_t = 0|\omega_{z_{t-1}}, z_{t-1})P(z_t|d_t))^{\mathcal{I}(d_t = d_{t-1})} P(z_t|d_t)^{(1-\mathcal{I}(d_t = d_{t-1}))} \end{aligned}$$

The second term in Eq. 3.9 can be computed recursively using message passing starting at $t = T$ where $g(y_{t+1:T}|z_{t+1}, d_{t+1}, \vec{\pi}) = 1$ and moving backward

$$g(y_{t:T}|z_t, d_t, \vec{\pi}) = f(y_t|z_t) \sum_{z_{t+1}, d_{t+1}} P(z_{t+1}, d_{t+1}|z_t, d_t, \vec{\pi}) g(y_{t+1:T}|z_{t+1}, d_{t+1}\vec{\pi})$$

Once posteriors are computed, within a time step t, top-down sampling is used as:

$$d_t|d_{t-1}, z_{t-1}, y_{1:T}, \vec{\pi} \sim \textstyle\sum_{z_t} P(d_t, z_t|d_{t-1}, z_{t-1}, y_{1:T}, \vec{\pi})$$

Variable o_t is only sampled when $d_t = d_{t-1}$. Furthermore, z_t is sampled only when $o_t = 0$ or d_t is different from d_{t-1}, otherwise $z_t = z_{t-1}$.

$$o_t|z_{t-1}, y_{1:T}, \vec{\pi} \sim P(o_t|z_{t-1}, \omega_{z_{t-1}}) \sum_{z_t} P(z_t|o_t, \vec{\pi}) P(y_{t:T}|z_t, d_t, \vec{\pi}) \qquad d_t = d_{t-1}$$

$$z_t|d_t, o_t = 0, y_{1:T}, \vec{\pi} \sim P(z_t|d_t) P(y_{t:T}|z_t, d_t, \vec{\pi})$$

3.6 Experiments and Results

We demonstrate the utility of TSTM on physiologic heart rate (HR) and respiratory rate (RR) signals collected from 145 premature infants from our Stanford NICU dataset. Due to prematurity, these infants are extremely vulnerable, and complications during their stay in the NICU can adversely affect long term development. Our broader aim is to identify markers associated with and predictive of downstream health status.

Clinicians and alert systems implemented on ICU monitors utilize coarse information such as signal means (e.g., is the HR > 160 beats per minute) and discard the remaining signal. We use TSTM to infer whether there is information contained in the signal dynamics. Transient events can manifest in the HR or the RR signal independently (e.g., bradycardia is observed in the HR signal while apnea is primarily observed in the RR signal). Thus, for our experiments below, we run TSTM on each signal independently; simultaneous processing of both signals is also easily possible with TSTM using the vector autoregressive process observation model as described in section 3.3.2.

Roadmap: We first evaluate the goodness of fit of TSTM on each physiologic signal. We also evaluate the utility of TSTM for feature construction on a supervised learning task of *grade assignment*. We then perform a qualitative analysis of the learned words, topics and inferred infant-specific distributions for clinical relevance. Finally, we show an experimental comparison between TSTM and BP-AR-HMM.

3.6.1 Experimental Setup

For all our experiments, we preprocess the physiologic signals to remove a 40 minute moving average window; this allows us to capture characteristics only related to the dynamics of the signal (resulting HR signal shown in figure 3.1). For TSTM, we fix the number of topics, $D = 4$. Although this choice is flexible, for our dataset, we chose this based on clinical bias. We identify four clinically meaningful topics: *Lung* for primarily lung related complications; *Head* for head related (neurological) complications; *Multi* as the catch-all class for severe complications that often affect multiple organ systems; and *Healthy*. We set the truncation level L to 15. We experimented with different settings of the hyperparameters for TSTM. Of particular interest is the choice of κ and ρ which control word and topic length and can,

as a result, force the words to be longer or shorter.[3]For the reported experiments, α_l, γ and η were each set to 10, $\kappa = 25$ and $\rho = 20$. Similar to [Fox *et al.*, 2009], we specify the priors on the observation model parameters A and Σ as a Matrix Normal Inverse-Wishart of $\mathcal{N}(0, 10 * I_p)$ and $IW(S_0, 3)$; S_0 is set to 0.75 times the empirical covariance of the first difference observations and p is the order of the autoregressive process. For all experiments with an AR observation model, we use this prior.

3.6.2 Quantitative Evaluation

Goodness-of-Fit

To evaluate the benefit of explicitly modeling heterogeneity between the time series, we compare the goodness-of-fit of TSTM with a switching-AR model on held out test data. Specifically, we compare with auto-regressive hidden Markov models (AR-HMM). AR-HMMs are HMMs where the observation model used are autoregressive processes [Poritz, 2009]. AR-HMMs, like other switching Markov models we discussed previously, assume that the data is generated i.i.d. from a single class model. Thus, an improved goodness-of-fit with TSTM will illustrate the benefit of modeling series-specific variability.

For the choice of the observation model, we experiment with both first (AR(1)-HMM) and second order (AR(2)-HMM) autoregressive processes. In figure 3.4, we illustrate the protocol for this experiment. As shown in figure 3.4a, to generate the test set on which we evaluate the fit, for each series, we hold out a sample of 4-hour blocks comprising 20% of the series. We keep the remaining as training data.

To evaluate the test log-likelihood, we average over three separate Gibbs chain for each model (see figure 3.4b). First, we run each chain for 2000 iterations on the training data. Chains for both models appear to mix by the $2000th$ iteration. Each chain is initialized by sampling model parameters (words, topics and topic-transition matrices for TSTM, and words and transition matrix for AR-HMM) from the prior. The training is done in an unsupervised way; the number of topics is initialized as $D = 4$ but no supervision is given regarding which infant contains which topics.

[3]We tested a few other settings of these hyperparameters. We qualitatively evaluated the word histograms (e.g., shown in figure 3.7a) derived from the series segmentations for whether the infants clustered based on their illness severity i.e. whether healthy infants have similar word profiles and profiles that are different from the unhealthy infants. This separation held consistently for our settings suggesting that with regard to discovery, the results are not sensitive to the specific choice of parameters. Since our primary goal is discovery, we ran all our experiments with only one setting of these hyperparameters.

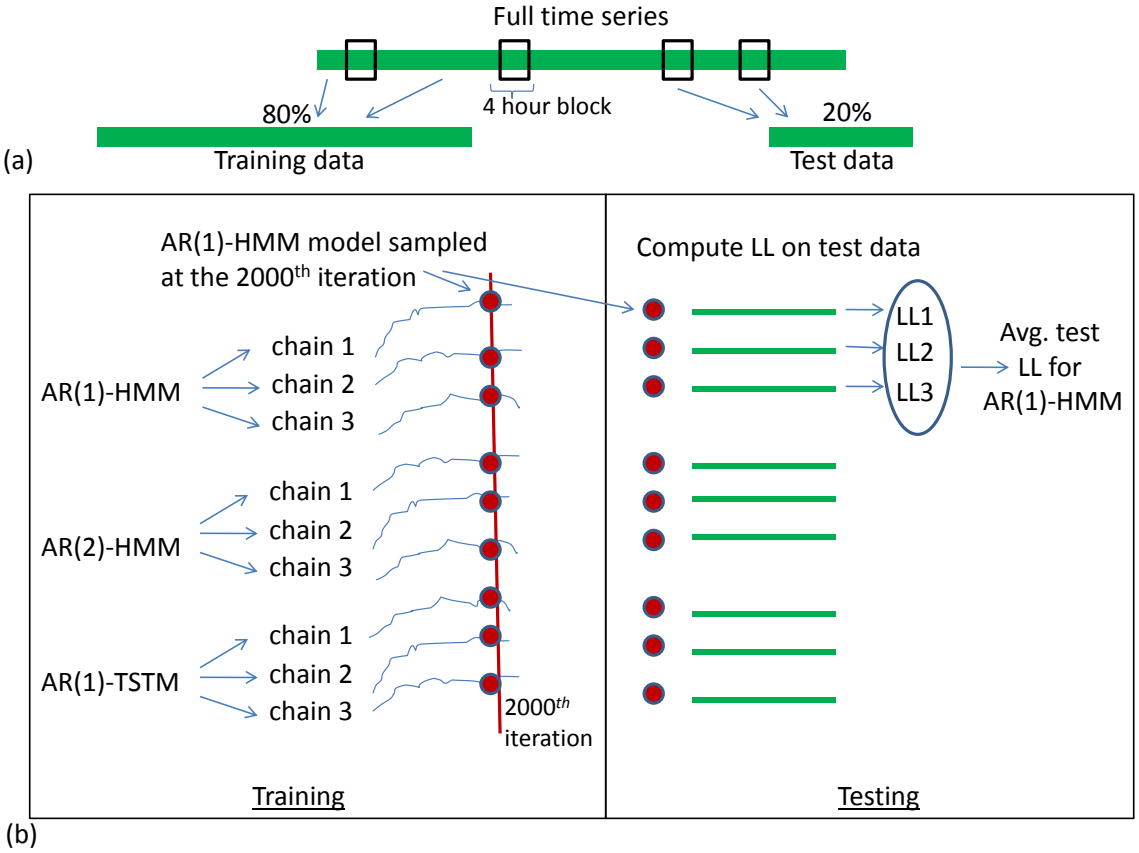


Figure 3.4: Experimental protocol for the evaluation of goodness-of-fit, a) the procedure for splitting each series into the train and test set, b) the pipeline for evaluating goodness of fit on the data.

We choose the model parameters sampled at the 2000th iteration as the model for which we evaluate the goodness-of-fit.[4] Each AR-HMM chain was initialized to have the same number of AR features as that inferred by the corresponding TSTM chain at the 2000th iteration. Test log-likelihood is computed with the forward-backward algorithm on the test sequences with model parameters fixed from the 2000th iteration of each Gibbs chain on the training data. Test log-likelihoods, averaged over 3 chains, for TSTM with an AR(1) observation model and an AR(1)-HMM are -1.425e+5 and -2.512e+5 respectively

[4]The 2000th iteration for each chain was chosen arbitrarily. Alternately, one may also choose more than one iteration and average over the iterations. This would require running inference on the test data with the model at each of these different iterations.

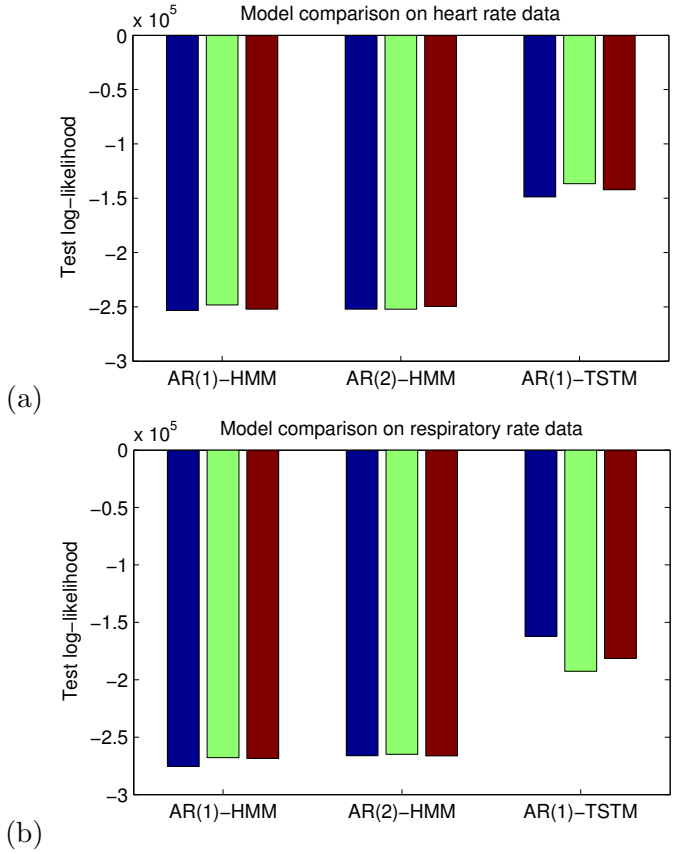


Figure 3.5: Test log-likelihood from three separate Gibbs chains for the AR(1)-HMM, AR(2)-HMM, and TSTM with an AR(1) observation model evaluated on a) the heart rate data (top), b) the respiratory rate data (bottom).

for the HR signal, and -1.787e+5 and -2.706e+5 for the RR signal. These results are also shown in figure 3.5a and figure 3.5b respectively. The significant gains in test log-likelihood using TSTM suggest that explicitly modeling heterogeneity between series is beneficial and that the topics and words generalize well to held-out data. In figure 3.5, we also see that the AR(2) observation model, albeit a more complex model than AR(1), does not benefit test log-likelihood. Thus, for all experiments that follow, we use an AR(1) observation model.

Feature Derivation

Deriving features is a common task one needs to tackle when using high-dimensional data (such as the physiologic signals) for supervised machine learning problems. We evaluate the usefulness of features obtained from the physiologic signals for the task of grade assignment. Grades $G_{1:N}$, representing an infant's health, are assigned to each infant based on his final outcome, as identified retrospectively by a clinician. Grade 0 is assigned to infants with no complications; grade 1 to isolated minor complications frequently associated with prematurity; grade 2 to multiple minor complications; and grades $3-5$ to major complications from low to severe grades. Features derived from any given model are used for predicting an infant's disease grade by combining these features with a rank support vector machine [Joachims, 2006]. The rank score for a ranking H is:

$$\text{rankscore}_H = \Sigma_{n=1}^{N}\Sigma_{m=1}^{N}\mathcal{I}(H(n) > H(m))(G_n - G_m)$$

We compute features from TSTM as follows. We run three Gibbs chains with TSTM (using unsupervised training as described in the goodness-of-fit experiment above) on the full data set. The features for each infant are derived as the frequency of each topic at the 2000th iteration of a Gibbs chain normalized by the length of the data sequence.

To report ranking test accuracy, for the set of 145 infants, we generate 20 random folds with $50-50$ train/test split and average performance over all folds. For each fold, the accuracy is computed as a percentage of the maximum achievable score for that test split. The SVM tradeoff parameter C for each model was set using cross-validation with features generated from the first Gibbs chain on 3 randomly sampled folds. Due to small sample size, we do not experiment with the choice of kernel and use the default choice of a linear kernel for all our experiments.

For comparison with other feature extraction methods from time series data, existing approaches can be divided into two broad classes: techniques in the frequency-domain and the time-domain [Shumway, 1988; Keogh *et al.*, 2000]. Frequency analysis using the discrete fourier transform is one of the most commonly used techniques for time series data analysis [Keogh *et al.*, 2000]. The frequencies of the resulting FFT coefficients span $1/v$ for $v \in \{1, \cdots, T\}$, which results in a large feature set. Traditionally, the large feature set size is not a concern in the presence of enough data. However, in our application, as is in most clinical applications, labeled data is often scarce. We experiment with using the raw features

within the rank SVM. Based on preliminary data analysis, we also compute transformed features by summing coefficients corresponding to time periods in increments of 4 minutes. This non-linear binning of features dramatically improves performance for HR data from near random to 63.5%. In the time domain, we compare with features derived from the AR(1)-HMM model; for this, we run three Gibbs chains on the full dataset (as described above). Features for each infant series are computed as the normalized proportion of words at the 2000th iteration of a Gibbs chain. Performance is reported as the average over all three chains. To get an assessment of the information contained in the dynamics compared to the signal mean (a simple measure usually used in standard care), we also grade infants based on deviation of their signal mean from the normal mean (normal specified as 125 beats/min for HR and 55 breaths/min for RR). We call this approach the clinical norm.

In table 1, we report results for all four methods. TSTM features yield higher performance for both the heart rate and respiratory rate signals compared to those derived from FFTs, AR-HMMs or the clinical norm (although statistical significance is not reached for difference between AR-HMM and TSTM performance). This suggests that the inferred topic proportions provide a useful feature representation scheme that can be employed within supervised objectives as an alternative or in addition to these existing methods.

Model	Heart Rate	Resp. Rate
TSTM	**74.45**%	**75.48**%
FFTs	63.5%	67.69%
AR-HMMs	71.29%	72.68%
Clinical norm	61.37%	68.93%

Table 3.2: Evaluating features from unsupervised training of TSTM.

Comparison to BP-AR-HMM

In figure 3.6a, we show an example run of the BP-AR-HMM model on the heart rate signal for a randomly selected set of 30 infants. For comparison, inference using TSTM on data from the same infants is shown in figure 3.7a. For the BP-AR-HMM run, we used a prior of Gamma$(1, 1)$ on α, the hyperparamter that controls the number of new features generated and Gamma$(100, 1)$ on κ, the self-transition parameter. The gamma proposals used $\sigma_\gamma^2 = 1$ and $\sigma_\kappa^2 = 200$. We refer the reader to [Fox *et al.*, 2009] for the definitions of these parameters.

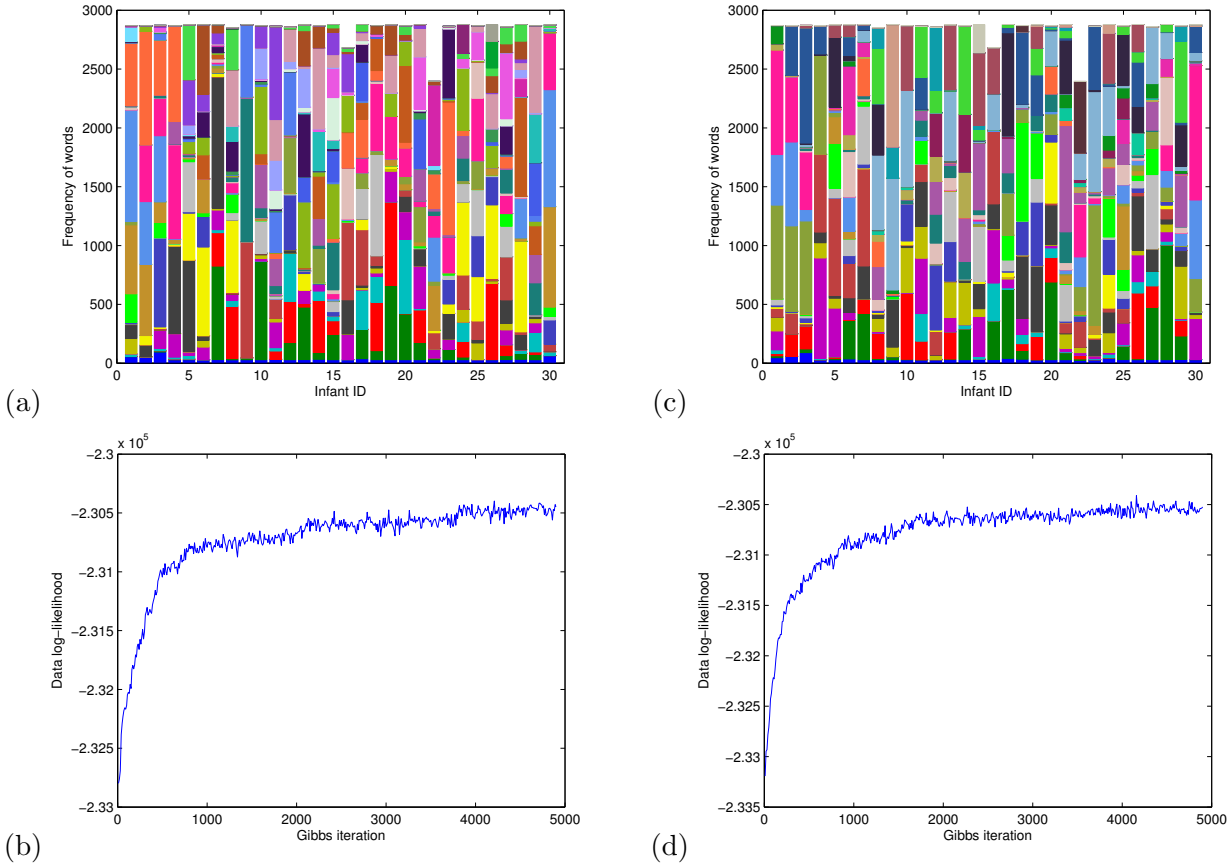


Figure 3.6: a) & c) Inferred word distributions from the heart rate signal for 30 infants during their first four days at the NICU with the BP-AR-HMM for two different initializations (initialization setting described in the text); distinct colors correspond to distinct words, b)& d) Corresponding data log-likelihood of the Gibbs chain for the first 5000 iterations.

For the observation model parameter priors, we use the same specification as that for the TSTM and AR-HMM experiments. At initialization, each series was segmented into five contiguous blocks, with feature labels unique to that sequence. BP-AR-HMM is sampling in the space of binary matrices; as a result, the birth-death sampler in [Fox *et al.*, 2009] takes much longer to mix compared to our block-Gibbs sampler. Due to the computational requirements of BP-AR-HMM, each chain is run on only the first four days of data; one such run takes approximately 34 hours. In the figure 3.6a, we show results from the 5000th iteration of a Gibbs chain. Each unique color corresponds to a distinct AR-feature (word) and the bar graph displays the distribution of words for each series.

In comparing the inferred word distributions from the BP-AR-HMM (figure 3.6a) with that from TSTM (figure 3.7a), we see that the inferred word distributions from the BP-AR-HMM are not as clearly amenable to clinical interpretation. Specifically, most series have features that are not shared with other series. To investigate whether the fragmentation was due to lack of mixing, in figure 3.6b, we plot the data log-likelihood as $\log \prod_i \sum_{Z^i} P(Y^i, Z^i | \pi^i, \vec{\theta})$; the chain appears to have mixed by the 5000th iteration. To encourage more sharing across the time series with the BP-AR-HMM, we ran several different Gibbs chains with different variants of the hyperparameters but the results were visually similar or even more fragmented. We show another example run in figure 3.6c where we use the prior for κ as Gamma$(200, 1)$ and at initialization, segment each series into only two blocks instead of five.[5]

This behavior is not entirely surprising, since the notion of individual series variation in BP-AR-HMM is quite different from that in TSTM: TSTM encourages sharing by having all series utilize the same set of topics, but to different extents; by contrast, BP-AR-HMMs uses the Beta prior for generation of series which posits that each series samples some features that are shared and others that are explicitly series-specific. The abundance of unique features makes comparison between series based on these features difficult.

In order to perform a quantitative comparison with the BP-AR-HMM features, since we only have a small set of 30 samples and therefore, grading does not make sense, we train an SVM classifier using leave-one-out cross-validation, to distinguish *Healthy* vs *Not healthy*, using the labels shown at the bottom of figure 3.7a. For the BP-AR-HMM, we compute frequencies of words extracted from the final iteration of the Gibbs chain shown in figure 3.6a as features. For TSTM, we run a Gibbs chain on the data for 30 infants without any supervision. The topic proportions at the 2000th iteration are used as features within the SVM. The inferred 4-topic proportions from TSTM yields an accuracy of 80% for HR and 60% for RR data. In contrast, BP-AR-HMM word proportions used as features yields a lower performance of 70% and 53%. Thus, the fragmentation of the data across multiple individual features hurts BP-AR-HMM's performance.

[5]The colors in these figures are generated randomly for each run so the colors are not comparable between figure 3.6a and figure 3.6c.

3.6.3 Qualitative Evaluation

We now analyze the words and topics inferred from TSTM in more detail. We focus on the model inferred from the heart rate signal.

Partially-supervised training: We experiment with the partially-supervised training regime of labeled LDA [Ramage *et al.*, 2009], which has the advantage of biasing the topics into categories that are coherent and more easily interpreted. During training, we constrain infant-specific transition matrices to *not* have topics corresponding to complications that they did not show symptoms for. This type of negative evidence imposes minimal bias, particularly relevant in clinical tasks, because of the uncertainty associated with the diagnosis of the onset and severity of the complication. For each infant in a randomly chosen subset of 30 infants, we assign a vector λ_n of length D, where we have a 0 at index i when this infant is known not to have complications related to the ith category. All infants are marked to allow having the healthy topic, representing the assumption that there may be some fraction of their time in the NICU during which they have recovered and are healthy. Each row of the infant-specific transition matrix is generated as:

$$\pi_n^i \sim \text{Dir}\left(\alpha_l \frac{\pi_g^i \otimes \lambda_n}{< \pi_g^i, \lambda_n >}\right) \qquad \lambda_n(i) = 1 \tag{3.10}$$

where $\otimes$ denotes the element-wise vector product. Under this regime, we run a Gibbs chain (G1) for the 30 infants. Next, we fix the words and topic distributions $\phi_{1:D}$ to that of the $2000th$ Gibbs iteration (as discussed in previous experiments) and run inference on our entire set of 145 infants. Here, no supervision is given; that is, both π_g and π_n are initialized from the prior and are left unconstrained during the inference process (using block Gibbs). We run a Gibbs chain to 400 iterations. Given the words and topics, the block-Gibbs sampler appears to mix within 200 iterations.

Qualitative analysis: In figure 3.7a, we show 30 randomly selected infants from this test set at the $400th$ iteration from chain 1. These infants are not the same as the infants used in the training set. In panel 3(a), we plot the word distribution for days 1,2 (top) and days 7,8 (bottom). Infants with no complications are shown as red squares at the bottom of this panel. In panel 3(b), we plot the degree to which a word is associated with each of the four topics.

First, we examine the inferred topic posteriors to track the clinical evolution of three

sample infants 2, 16 and 23 chosen to be illustrative of different trajectories of the word distributions over time. In figure 3.7c, the bold line shows the smoothed posterior over the infant being healthy over time. To compute this posterior, for each of the three infants we run 30 test Gibbs chains with words and topics fixed from the final iteration of G1 (described above). For each infant, at time t within its sequence, we compute h_t as the proportion of times latent state d_t = Healthy from the final iteration of all 30 chains. We smooth h_t using an 8 hour window around t. Thus, a posterior value of 1 at t implies the infant only expressed words associated with the Healthy topic within an 8 hour window of t.

Infant 2 (I2) was born with a heart defect (small VSD). I2's heart condition was treated on day 4. On day 7, her state started to resolve significantly, and on day 8 her ventilator settings were minimal and she was taken off the drug. Her empirical evolution closely tracks her medical history; in particular, her state continually improves after day 4. Infant 16 was a healthier preemie with few complications of prematurity and was discharged on day 4. Infant 23, on the other hand, got progressively sicker and eventually died on day 4. The figure shows that their inferred posterior prediction closely tracks their medical history as well.

Next, we analyze the words and word histograms. Loosely interpreting, words with AR parameter $a > 1$ represent heart rate accelerations (e.g., word 8 shown in gray), words where a is positive and close to 0 represent periods with significantly lower dynamic range (e.g., word 2 shown in purple) and words with large V represent higher dynamic range or high entropy. The word frequencies vary greatly across infants. Respiratory distress (RDS), a common complication of prematurity, usually resolves within the first few days as the infant stabilizes and is transitioned to room air. This is reflected by the decrease in relative proportion of word 2, only associated with the Lung topic (as seen in figure 3.7b). Exceptions to this are infants 3 and 30, both of whom have chronic lung problems. Overall, the inferred word histograms highlights separability between healthy and other infants based on the word mixing proportions, suggesting different dynamics profiles for these two populations. Words 3, 9 and 10, associated primarily with the healthy topic, occur more frequently in infants with no complications. These three words also have the highest V^k values suggesting entropy as a signature for health in neonates. Thus, we developed a new risk stratification score [Saria *et al.*, 2010], that predicts based on data from the first three hours of life, infants at risk for major complications. We describe this score in detail in

Chapter 5.

3.7 Discussion and Future work

The primary contribution of this chapter is a new class of models for time-series data that emphasizes the modeling of instance specific variability while discovering population level characteristics. Unlike BP-AR-HMM, its closest counterpart, TSTM has a mechanism for modeling high-level topics that hierarchically capture structure in collections of words. For a knowledge discovery task, modeling higher-level structure has several advantages. First, it can help discover novel semantic structure in the data such as the degree to which two topics (e.g., diseases) share common words (e.g., physiologic traits). It also gives the used finer control over the types of features extracted from the data; for example, by using TSTM within a partially supervised setting, emphasis can be placed on discovering features that identify specific disease pairs.

We demonstrate the use of TSTM in a novel and useful application of modeling heterogeneous patient populations over time. We believe that TSTM provides a significant departure from current practices and a flexible tool for exploratory time series data analysis in novel domains. Furthermore, learned topic or word distributions can serve as features within supervised tasks. We demonstrated the utility of TSTMs on medical time series, but the framework is broadly applicable to other time-series applications such as financial or human-activity data [Liao *et al.*, 2007].

Several extensions of TSTM could yield additional domain insight. In particular, modeling individual topic distributions as evolving over time (analogous to [Wang *et al.*, 2008]) should highlight how the characteristics of the expressed temporal signatures vary as diseases evolve over longer periods of time. Modeling the data as the composition of repeating signatures expressed at varying temporal granularity (seconds versus minutes versus hours) would highlight the granularity at which diseases alter measured physiology. We leave these next steps for future work. In the next chapter, we extend the notion of words from dynamics related signtures to richer shape related signatures.

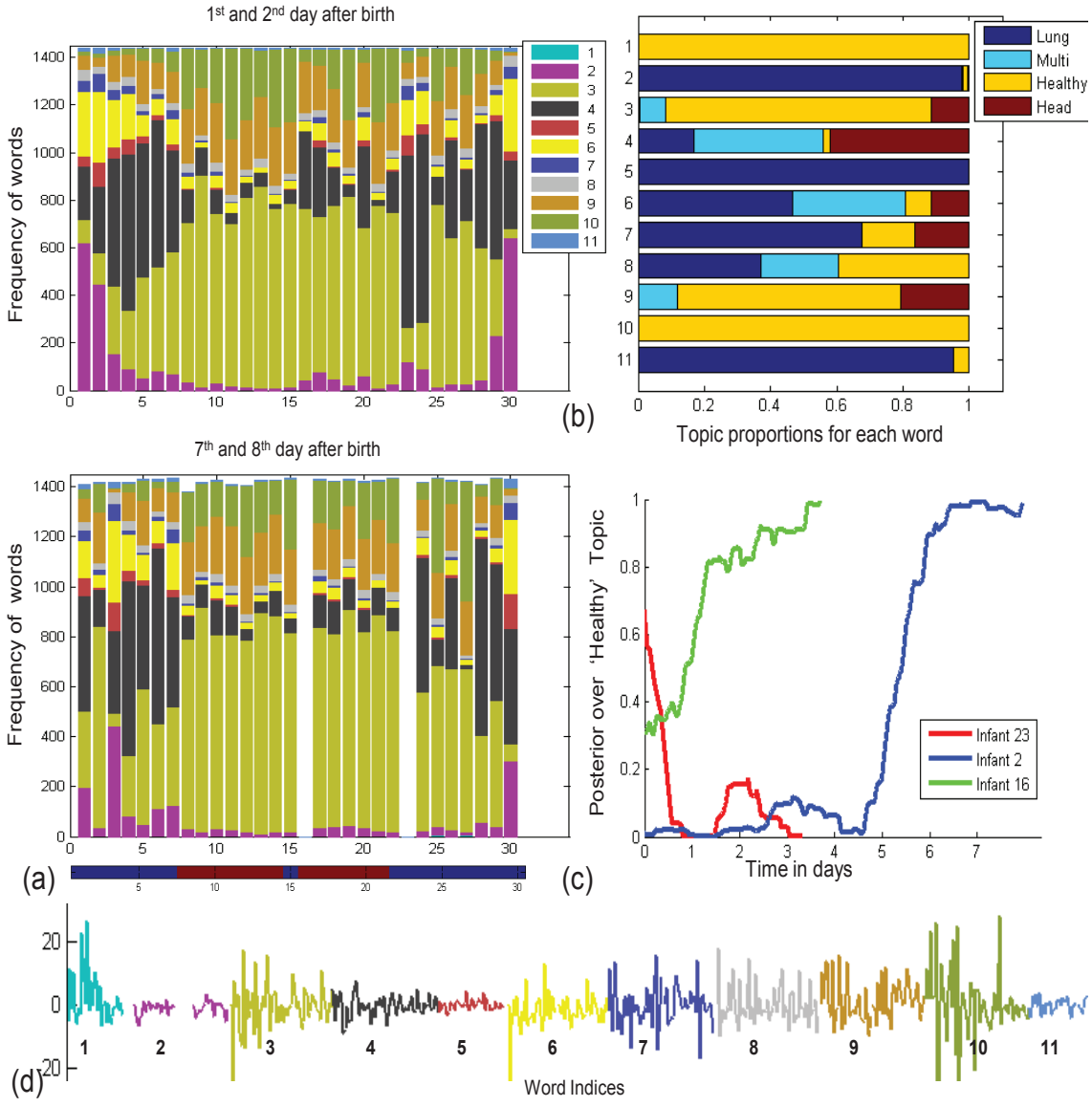


Figure 3.7: (a) Inferred word distributions for the heart rate data for 30 infants during their stay at the NICU. At the bottom of the word panel, infants marked with red squares have no complications, (b) distribution over disease topic given words for the population, (c) posterior over latent state, *Healthy*, (d) examples of inferred features extracted from the data.

Chapter 4

Discovering Shape Signatures in Physiologic data

In addition to dynamics signatures, continuous time series data may often comprise or contain repeated *motifs* — patterns that have similar shape, and yet exhibit nontrivial variability. Identifying these motifs, even in the presence of variation, is an important subtask in both unsupervised knowledge discovery and constructing useful features for discriminative tasks. This chapter addresses this task using a probabilistic framework that models generation of data as switching between a random walk state and states that generate motifs. A motif is generated from a continuous shape template that can undergo non-linear transformations such as temporal warping and additive noise. We propose an unsupervised algorithm that simultaneously discovers both the set of canonical shape templates and a template-specific model of variability manifested in the data. Experimental results on both physiologic signals from infants, as well as two other challenging real-world domains, demonstrate that our model is able to recover templates in data where repeated instances show large variability. The recovered templates provide higher classification accuracy and coverage when compared to those from alternatives such as random projection based methods and simpler generative models that do not model variability. In analyzing physiological signals from infants in the ICU, we discover both known signatures as well as novel physiomarkers.

4.1 Introduction

Continuous-valued time series data in several domains (e.g., pose tracking, finance, patient monitoring) often contain *motifs* — segments with similar structure that repeat within and across different series. For example, in physiologic signals, recognizable shapes such as bradycardia and apnea are known to precede severe complications such as infection. In trajectories of people at an airport, we might see repeated motifs in a person checking in at the ticket counter, stopping to buy food, etc. In pose tracking, we might see characteristic patterns such as bending down, sitting, kicking, etc. Discovering these repeated segments can provide primitives that are useful for domain understanding and as higher-level, meaningful features that can be used to segment time series or discriminate among time series data from different groups.

In many domains, different instances of the same motif can be structurally similar but vary greatly in terms of pointwise distance [Höppner, 2002]. For example, the temporal position profile of the body in a front kick can vary greatly, depending on how quickly the leg is raised, the extent to which it is raised and then how quickly it is brought back to position. Yet, these profiles are structurally similar, and different from that of a roundhouse kick. Bradycardia and apnea are also known to manifest significant variation in both amplitude and temporal duration. Our goal is to deal with the unsupervised discovery of these *deformable* motifs in continuous time series data.

Much work has been done on the problem of motif detection in continuous time series data. One very popular and successful approach is the work of Keogh and colleagues (e.g., [Mueen *et al.*, 2009]), in which a motif is defined via a pair of windows of the same length that are closely matched in terms of Euclidean distance. Such pairs are identified via a sliding window approach followed by random projections to identify highly similar pairs that have not been previously identified. However, this method is not geared towards finding motifs that can exhibit significant deformation. Another line of work tries to find regions of high density in the space of all subsequences via clustering; see Oates [2002]; Denton [2005] and more recently Minnen et al. [2007]. These works define a motif as a vector of means and variances over the length of the window, a representation that also is not geared to capturing deformable motifs. Of these methods, only the work of Minnen et al. [2007] addresses deformation, using dynamic time warping to measure warped distance. However, motifs often exhibit structured transformations, where the warp changes gradually

over time. As we show in our results, encoding this bias greatly improves performance. The work of Listgarten et al. [2005]; Kim et al. [2006] focuses on developing a probabilistic model for aligning sequences that exhibit variability. However, these methods rely on having a segmentation of the time series into corresponding motifs. This assumption allows them to impose relatively few constraints on the model, rendering them highly under-constrained in our unsupervised setting.

Here, we propose a method, which we call CSTM (Continuous Shape Template Model), that is specifically targeted to the task of unsupervised discovery of deformable motifs in continuous time series data. CSTM seeks to explain the entire data in terms of repeated, warped motifs interspersed with non-repeating segments. In particular, we define a hidden, segmental Markov model in which each state either generates a motif or samples from a non-repeating random walk (NRW). The individual motifs are represented by smooth continuous functions that are subject to non-linear warp and scale transformations. Our warp model is inspired by Listgarten et al. [2005], but utilizes a significantly more constrained version, more suited to our task. We learn both the motifs and their allowed warps in an unsupervised way from unsegmented time series data. We demonstrate the applicability of CSTM to three distinct real-world domains and show that it achieves considerably better performance than previous methods, which were not tailored to this task.

4.2 Generative Model

The CSTM model assumes that the observed time series is generated by switching between a state that generates non-repeating segments and states that generate repeating (structurally similar) segments or *motifs*. Motifs are generated as samples from a shape template that can undergo non-linear transformations such as shrinkage, amplification or local shifts. The transformations applied at each observed time t for a sequence are tracked via latent states, the distribution over which is inferred. Simultaneously, the canonical shape template and the likelihood of possible transformations for each template are learned from the data. The random-walk state generates trajectory data without long-term memory. Thus, these segments lack repetitive structural patterns. Below, we describe more formally the components of the CSTM generative model. In Table 4.1, we summarize the notation used for each component of the model and other notation used frequently through the rest of the chapter.

Symbol	Description
y_t	Observation at time t
κ_t	Index of the template used at time t
ρ_t	Position within the template at time t
ω_t	Temporal warp applied at time t
ϕ_t	Scale tranformation applied at time t
ν_t	Additive noise at time t
z_t	Vector $\{\rho_t, \phi_t, \nu_t\}$ of transformations at time t
s^k	kth template, length of template is L^k
π_ω^k	Warp transition matrix for kth template
π_ϕ^k	Scale transition matrix for kth template
$\mathcal{T}$	Transition matrix for transitions between templates and NRW
σ	The variance parameter for the NRW state
$\dot{\sigma}$	The variance parameter for all shape template states
λ	Controls proportion of data explained by NRW vs. the shape template states
$\mathcal{I}(E)$	Indicator function where E is the event

Table 4.1: Notation for the generative process of CSTM and other frequently used notation in this chapter.

4.2.1 Canonical Shape Templates (CST)

Our main goal with the CSTM is to uncover the prototypical 'shape templates' that reflect the shapes of repeated occurrences in our time series Y. Even though the observed series Y is measured at discrete times, CSTM models the underlying shape over the continuous range of the length of the template.

Each shape template, indexed by k, is represented as a continuous function $s^k(l)$ where $l \in (0, L^k]$ and L^k is the length of the kth template. Although the choice of function class for s^k is flexible, a parameterization that encodes the property of motifs expected to be present in given data will yield better results. In many domains, motifs appear as smooth functions. A possible representation might be an L_2-regularized Markovian model i.e. a chain where the difference between neighboring values receives a squared penalty. However, this representation penalizes smooth functions with a trendline (non-zero gradient) more than those that are flat, a bias not always justified. A promising

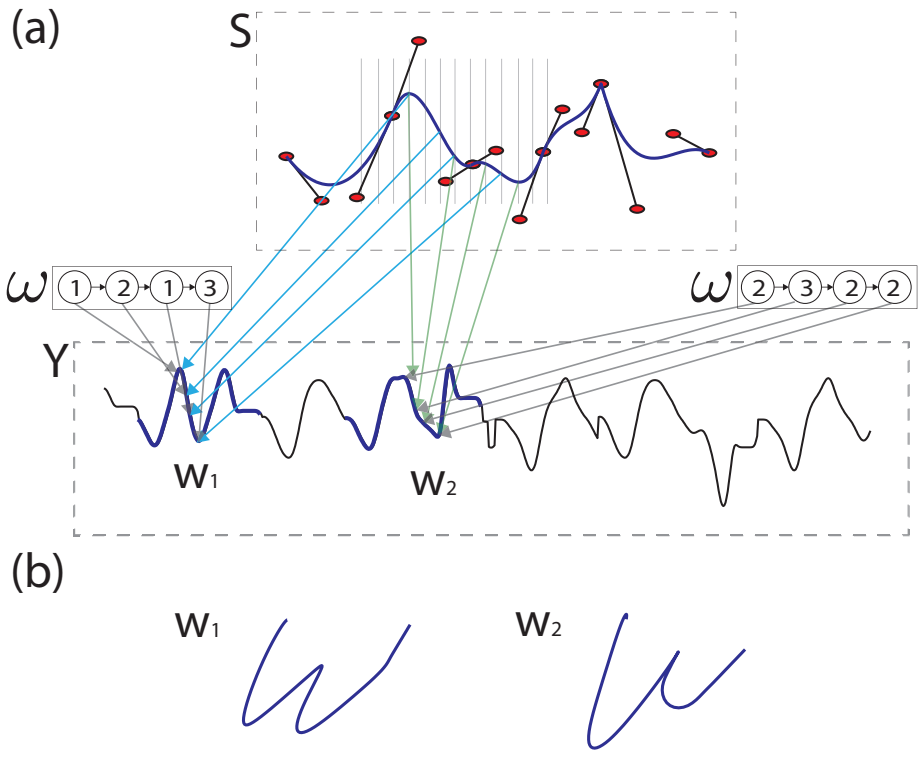


Figure 4.1: a) The template S shows the canonical shape for the pen-tip velocity along the x-dimension and a piecewise Bézier fit to the signal. The generation of two different transformed versions of the template are shown; for simplicity, we assume only a temporal warp is used and ω tracks the warp at each time, b) The resulting character 'w' generated by integrating velocities along both the x and y dimension.

alternative is piecewise Bézier splines [Gallier, 1999]. Shape templates of varying complexity are intuitively represented by using fewer or more pieces. For our purpose, it suffices to present the mathematics for the case of piecewise third order Bézier curves over two dimensions, where the first dimension is the time t and the second diemsnion is the signal value.

A third order Bézier curve is parameterized by four points $p_i \in \mathcal{R}^2$ for $i \in 0, \cdots, 3$. Control points p_0 and p_3 are the start and end of each curve piece in the template and

shared between adjacent pieces (see figure 4.1). The intermediate control points p_1 and p_2 control the tangent directions of the curve at the start and end, as well as the interpolation shape. Between end points, $\tau \in [0, 1]$ controls the position on the curve, and each piece of the curve is interpolated as

$$f(\tau) = \sum_{i=0}^{3} \begin{pmatrix} 3 \\ i \end{pmatrix} (1-\tau)^{3-i} \tau^i p_i \tag{4.1}$$

For higher-dimensional signals in $\mathcal{R}^n$, $p_i \in \mathcal{R}^{n+1}$. We use this representation in all our experiments with real-world data. Also, even though only $\mathcal{C}^0$ continuity is imposed here, it is possible to impose arbitrary continuity within this framework of piecewise Bézier curves if such additional bias is relevant.

4.2.2 CST Deformation Model

In most natural domains, every instantiation of a repeated pattern will not be exactly the same. Consider the case of written letters in figure 4.1b: the w's have similar structure as they are both easily interpreted as the same symbol, but the handwriting differs leading to noticeable differences between the two instances.

These differences can be changes in scale (e.g., the second half of the second instance of w is smaller than the rest of the character), changes in speed (e.g., one person can do the jumping jack faster than the other leading to a temporally squeezed pose profile), or random insertions (e.g., a heart rate monitor may given an erroneous reading for a second due to some interference). Loosely speaking, it is differences such as these that are unrelated to the overall structure that the transformation model seeks to capture.

Motifs are generated by non-uniform sampling and scaling of s^k. Temporal warp can be introduced by moving slowly or quickly through s^k. The allowable temporal warps are specified as an ordered set $\{w_1, \ldots, w_n\}$ of time increments that determines the rate at which we advance through s^k. A template-specific warp transition matrix π_ω^k specifies the probability of transitions between warp states. To generate a series $y_1, \cdots, y_T$, let $\omega_t \in \{w_1, \ldots, w_n\}$ be the random variable tracking the warp and ρ_t be the position within the template s^k at time t. Then, y_{t+1} would be generated from the value $s^k(\rho_{t+1})$ where $\rho_{t+1} = \rho_t + \omega_{t+1}$ and $\omega_{t+1} \sim \pi_\omega^k(\omega_t)$. (For all our experiments, the allowable warps are $\{1, 2, 3\}\delta t$ where δt is the sampling rate; this posits that the longest sequence from s^k is

at most three times the shortest sequence sampled from it.) In figure 4.1a, we show an illustration of the temporal warp to generate two different instances from the same symbol.

We might also want to model scale deformations. Analogously, the set of allowable scaling coefficients are maintained as the set $\{c_1, \ldots, c_n\}$. Let $\phi_{t+1} \in \{c_1, \ldots, c_n\}$ be the sampled scale value at time $t+1$, sampled from the scale transition matrix π_ϕ^k. Thus, the observation y_{t+1} would be generated around the value $\phi_{t+1} s^k(\rho_{t+1})$, a scaled version of the value of the motif at ρ_{t+1}, where $\phi_{t+1} \sim \pi_\phi^k(\phi_t)$. Finally, an additive noise value $\nu_{t+1} \sim \mathcal{N}(0, \dot{\sigma})$ models small shifts. The parameter $\dot{\sigma}$ is shared across all templates.

In summary, we use $z_t = \{\rho_t, \phi_t, \nu_t\}$ to represent the values of all transformations at time t and putting together all three possible deformations, we have:

$$y_{t+1} = \nu_{t+1} + \phi_{t+1} s^k(\rho_{t+1})$$

In many natural domains, motion models are often smooth due to inertia. For example, while kicking, as the person gets tired, he may decrease the pace at which he raises his leg. But, the decrease in his pace is likely to be smooth rather than transitioning between a very fast and very slow pace from one time step to another. One simple way to capture this bias is by constraining the scale and warp transition matrices to be band diagonal. Specifically, $\phi_\omega(w, w') = 0$ if $|w - w'| > b$ where $2b+1$ is the size of the band. (We set $b = 1$ for all our experiments.) Experimentally, we observe that in the absence of such a prior, the model is able to align random walk sequences to motif sequences by switching arbitrarily between transformation states, leading to noisy templates and poor performance. We also allow the warp and scale transition matrices to be template specific assuming that different shapes show different types and amounts of variations.

4.2.3 Non-repeating Random Walk (NRW)

We use the NRW model to capture data not generated from the templates (see also Denton [2005]). If this data has different noise characteristics, our task becomes simpler as the noise characteristics can help disambiguate between motif-generated segments and NRW segments. The generation of smooth series can be modeled using an autoregressive process. We use an $AR(1)$ process for our experiments where $y_t = \mathcal{N}(y_{t-1}, \sigma)$. We refer to the NRW model as the 0th template.

4.2.4 Template Transitions

Transitions between generating NRW data and motifs from the CSTs are modeled via a transition matrix, $\mathcal{T}$ of size $(K+1)\times(K+1)$ where the number of CSTs is K. The random variable κ_t tracks the template at t for an observed series. Transitions out of a template is not allowed until the end of the template is reached. Thus,

$$P(\kappa_t = \kappa_{t-1} | \rho_{t-1} + \omega_t < L^{\kappa_{t-1}}) = 1$$

$$P(\kappa_t \neq \kappa_{t-1} | \rho_{t-1} + \omega_t < L^{\kappa_{t-1}}) = 0$$

Otherwise, $\kappa_t \sim \mathcal{T}(\kappa_{t-1})$ as the current position has moved past the length of the template.

For $\mathcal{T}$, we fix the self-transition parameter for the NRW state as λ, a pre-specified input. Different settings of $0 < \lambda < 1$ allows control over the proportion of data assigned to motifs versus NRW. As λ increases, more of the data is explained by the NRW state and as a result, the recovered templates have lower variance.[1]

4.2.5 Summary of CSTM generative process

Below, we summarize the generative process at each time t:

$$\kappa_t \sim \mathcal{T}(\kappa_{t-1}, \rho_{t-1}) \tag{4.2}$$

$$\omega_t \sim \pi_\omega^{\kappa_t}(\omega_{t-1}) \tag{4.3}$$

$$\rho_t = \begin{cases} \min(\rho_{t-1} + \omega_t, L^{\kappa_t}), & \text{if } \kappa_t = \kappa_{t-1} \quad (4.4) \\ 1, & \text{if } \kappa_t \neq \kappa_{t-1} \quad (4.5) \end{cases}$$

$$\phi_t \sim \pi_\phi^{\kappa_t}(\phi_t) \tag{4.6}$$

$$\nu_t \sim \mathcal{N}(0, \dot{\sigma}) \tag{4.7}$$

$$y_t = \nu_t + \phi_t s^{\kappa_t}(\rho_t) \tag{4.8}$$

[1] Learning λ while simultaneously learning the remaining parameters leads to degenerate results where all points end up in the NRW state with learned $\lambda = 1$.

4.3 Learning the model

For learning the model, our goal is to find the canonical shape templates, their template-specific transformation models, the NRW model, the template transition matrix and the latent states $(\kappa_{1:T}, z_{1:T})$ for the observed series from the data. This can be done by optimizing the log-likelihood with respect to each of these variables. However, the proposed log-likelihood function does not have a closed form solution and is non-convex.

We use hard Expectation Maximization (EM), a block coordinate ascent algorithm, which approximately optimizes the joint likelihood of the data and the latent variables. We opted for hard assignments in the E-step as it is considerably computationally faster when coupled with pruning.

Let Θ be an arbitrary model and $P_\Theta(y_{1:T}, \kappa_{1:T}, z_{1:T})$ denote the corresponding joint likelihood. Then, hard EM approximately solves the optimization problem:

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}} \max_{\kappa_{1:T}, z_{1:T}} P_\Theta(y_{1:T}, \kappa_{1:T}, z_{1:T})$$

Until convergence, it performs

- E-step: the maximum likelihood assignment to all latent variables given a model specification

$$\kappa_{1:T}, z_{1:T} := \underset{\kappa_{1:T}, z_{1:T}}{\operatorname{argmax}} P_\Theta(y_{1:T}, \kappa_{1:T}, z_{1:T})$$

- M-step: the maximum likelihood assignment to all model parameters given an assignment to the latest variables

$$\Theta := \underset{\Theta}{\operatorname{argmax}} P_\Theta(y_{1:T}, \kappa_{1:T}, z_{1:T})$$

This block coordinate procedure guarantees convergence to a local optimum. Below, we describe each step of the learning algorithm in more detail.

4.3.1 E-step

The E-step relies on known algorithms in a relatively straightforward manner. Given the series $y_{1:T}$ and the model parameters Θ from the previous M-step iteration, in the E-step, we

compute assignments to the latent variables $\{\kappa_{1:T}, z_{1:T}\}$ using approximate Viterbi decoding [Viterbi, 1967].

The approximation done in our Viterbi is to maintain only the high-probability template, warp and template-position hypotheses for each time point as this reduces the state space we must examine at the subsequent time point. At each time, we prune to maintain only the top B highest probability states. This approach of pruning and its variant of pruning with forward lookups has been used extensively in the speech processing community [Jelinek, 1997]. More specifically, we prune in the forward step as:

$$P(\kappa_t, z_t|y_{1:t}, \Theta) \propto P(y_t|\kappa_t, z_t)P(\kappa_t, z_t|\kappa_{t-1}, z_{t-1})\hat{P}(\kappa_{t-1}, z_{t-1}|y_{1:t-1}, \Theta)$$

$$\hat{P}(\kappa_t, z_t|y_{1:t}, \Theta) = P(\kappa_t, z_t|y_{1:t}, \Theta).*\vec{s}$$

$.*$ denotes element wise product. The vector $\vec{s}$ is of length $P(\kappa_t, z_t|y_{1:t}, \Theta)$. It contains as its elements 1 at indices where the corresponding state configuration κ_t, z_t is one of the top B highest probability states and 0 otherwise.

Exact inference in our model is feasible. If T is the length of the series, W and D are dimensions of the warp and scale transformation matrices respectively and K is the number of templates, then the length of the belief state at any given time is $W * D * K$ and the forward computation at each time t costs $\mathcal{O}(W^2 * D^2 * K^2)$. Thus, the total cost of exact inference is $\mathcal{O}(T * W^2 * D^2 * K^2)$. While exact inference is feasible, pruning significantly reduces the runtime of our algorithm. A few early tests indicated that pruning did not impact the final assignment as low probability states were seldom used in the final Viterbi assignments.[2] For our experiments, we maintain $K \times 20$ states.

4.3.2 M-step

In the M-step, given the data $y_{1:T}$ and the latent trace $\{\kappa_{1:T}, z_{1:T}\}$, the model parameters are optimized by taking the gradient of the penalized complete data log-likelihood with respect to each parameter. Below, we discuss the penalty for each component and the corresponding update equations.

[2] Although no theoretical guarantees exist for these pruning schemes, in practice they have been used extensively in the speech recognition community and shown to perform well.

Updating the transformation model, π_ω and π_ϕ

A Dirichlet prior, conjugate to the multinomial distribution, is used for each row of the transition matrices as penalty $\mathcal{P}_\omega$ and $\mathcal{P}_\phi$. In both cases, the prior matrix is constrained to be band-diagonal. As a result, the posterior matrices are also band-diagonal. The update is the straightforward MAP estimate for multinomials with a Dirichlet prior. If $\eta_{i,j}$ are the pseudo-counts from the prior, then the update for the posterior is:

$$\pi_\omega^k(i,j) \propto \eta_{i,j} + \sum_{t=1}^{T} \mathcal{I}(\kappa_t = k)\mathcal{I}(\omega_{t-1} = i)\mathcal{I}(\omega_t = j)$$

π_ϕ^k's are updated similarly.

For all our experiments, we set a weak prior favoring shorter warps: $\mathrm{Dir}(7,3,0)$, $\mathrm{Dir}(4,5,1)$ and $\mathrm{Dir}(0,7,3)$ for each of the rows of π_ω and π_ϕ given our setting of allowable warps. As always, the effect of the prior decreases with larger amounts of data. In our experiments, we found the recovered templates to be insensitive to the setting for a reasonable range.[3]

Updating the template transition matrix $\mathcal{T}$

The template transition matrix $\mathcal{T}$ is updated similar to the CST transformation matrices. A single hyperparameter $\dot{\eta}$ is used to control the strength of the prior. We set $\dot{\eta} = n/(K^2 L)$, where n is the total amount of observed data, L is the anticipated template length used in the initializations, and K is the pre-set number of templates. This is equivalent to assuming that the prior has the same strength as the data and is distributed uniformly across all shape templates. To update the transitions out of the NRW state,

$$\mathcal{T}_{0,k} = (1-\lambda)\frac{\dot{\eta} + \sum_{t=2}^{T}\mathcal{I}(\kappa_{t-1}=0)\mathcal{I}(\kappa_t = k)}{\dot{\eta}K + \sum_{t=2}^{T}\mathcal{I}(\kappa_{t-1}=0)\sum_{k'=1}^{K}\mathcal{I}(\kappa_t = k')}$$

Transitions are only allowed at the end of each template. Thus, to update transitions between shape templates,

$$\mathcal{T}_{k,k'} \propto \dot{\eta} + \sum_{t=2}^{T}\mathcal{I}(\kappa_{t-1}=k)\mathcal{I}(\kappa_t = k')\mathcal{I}(\rho_{t-1} = L^k)$$

[3] If the motifs exhibit large unstructured warp, the prior over the rows of the warp matrices can be initialized as a symmetric Dirichlet distribution. However, as seen in our experiments, we posit that in natural domains, having a structured prior improves recovery.

Fitting Shape Templates

Given the scaled and aligned segments of all observed time series assigned to any given shape template s^k, the smooth piecewise function can be fitted independently for each shape template. Thus, collecting terms from the log-likelihood relevant for fitting each template, we get:

$$\mathcal{L}_k = -\mathcal{P}_{s^k} - \sum_{t=1}^{T} \mathcal{I}(\kappa_t = k) \frac{(y_t - \phi_t s^k(\rho_t))^2}{\dot{\sigma}^2} \tag{4.9}$$

where $\mathcal{P}_{s^k}$ is a regularization for the kth shape template. A natural regularization for controlling model complexity is the BIC penalty [Schwarz, 1978] specified as $0.5 \log(N)\nu_k$, where ν_k is the number of Bézier pieces used and N is the number of samples assigned to the template.[4]

Piecewise Bézier curve fitting to chains has been studied extensively. We briefly review how we fit these curves to our data. $\mathcal{L}_k$ is not differentiable and non-convex; a series of hill-climbing moves are typically used to get to an optimum. We employ a variant which has been commercially deployed for large and diverse image collections [Diebel, 2008].

First we describe how to fit each curve-piece given the knots. Let the knots of the piecewise curve be $\{q_0 = 0, q_2, \cdots, q_{\nu_k} = L^k\}$. We can re-write $\mathcal{L}_k$ for any shape template k in terms of the indpendent contribution from piece as:

$$\mathcal{L}_k = -0.5 \log(N)\nu_k - \sum_{i=0}^{\nu_k - 1} \sum_{p=q_i}^{q_{i+1}} \sum_{t=1}^{T} \mathcal{I}(\kappa_t, k)\mathcal{I}(\rho_t, p) \frac{(y_t - \phi_t s^k(p))^2}{\dot{\sigma}^2} \tag{4.10}$$

Given the knots $[q_i, q_{i+1}]$, control points for a piece are estimated (independently from other pieces) using least squares regression with the data in that interval. Let M be the number of data samples specific to that interval, y_m their values and ρ_m their position within the template. Construct the data matrix $Y \in \mathcal{R}^{M \times 2}$ with each row as $[y_m \ \rho_m]$. Let τ_m be ρ_m interpolated between $[0, 1]$. Let Φ and Ψ be matrices for which the mth row is constructed as follows:

$$\Phi_m = \begin{bmatrix} (1-\tau_m)^3 & 3\tau_m(1-\tau_m)^2 & 3\tau_m^2(1-\tau_m) & \tau_m^3 \end{bmatrix},$$

[4] A modified BIC penalty of $\gamma(0.5 \log(N)\nu_k)$ can be used if further tuning is desired. Higher values of γ lead to smoother curves.

$$\Psi_m = \phi_m \Phi_m$$

where ϕ_m is the inferred scale for the mth data sample. Then $\mathbf{p}$, the matrix of control points, is computed as:

$$\mathbf{p} = \text{inv}(\Psi^T \Psi - \epsilon I)\Psi^T \vec{r}, \qquad \mathbf{p} \in \mathcal{R}^{4\times 2}, \qquad \Psi \in \mathcal{R}^{M\times 4}$$

$$\vec{r} = Y - \Psi \mathbf{p}_o^T, \qquad \vec{r} \in \mathcal{R}^{M\times 2}$$

$\epsilon \in \mathcal{R}_+$ is applied to make the matrix inversion non-singular. The control points that are shared with neighboring pieces are fixed and initialized as such in $\mathbf{p}_o$.

To optimize the placement of control point in $\mathcal{L}_k$, the algorithm defines break, swap and merge moves and applies them iteratively until no gain can be made in the objective value. A *break move* splits an interval $[t1, t2]$ into two equal length intervals. This move is accepted if the gain in likelihood is greater than $0.5\log(N)$, the penalty from adding a piece. A *swap move* adjusts the boundary between adjacent curve pieces. It swaps a subinterval containing the boundary point out of its current curve and into the adjacent curve. The swap move is accepted if the gain in likelihood isbpositive. The *merge move* merges two curve pieces into a single piece. This move is accepted if the loss in likelihood is less than $\log(N)$.

Updating σ and $\dot{\sigma}$

Given the assignments of the data to the NRW and the template states, and the fitted template functions, the variances σ and $\dot{\sigma}$ are computed easily.

$$\sigma = \sqrt{\frac{\sum_{t=2}^{T} \mathcal{I}(\kappa_t = 0)(y_t - y_{t-1})^2}{\sum_{t=2}^{T} \mathcal{I}(\kappa_t = 0)}}$$

$$\dot{\sigma} = \sqrt{\frac{\sum_{t=1}^{T} \mathcal{I}(\kappa_t \neq 0)(y_t - \phi_t s^{\kappa_t}(\rho_t))^2}{\sum_{t=1}^{T} \mathcal{I}(\kappa_t \neq 0)}}$$

4.3.3 Escaping local maxima

EM is known to get stuck in local maxima, having too many clusters in one part of the space and too few in another [Ueda *et al.*, 1998]. Split and merge steps can help escape these configurations by: a) splitting clusters that have high variance due to the assignment of a mixture of series, and b) merging similar clusters. At each such step, for each existing template k, 2-means clustering is run on the aligned segmentations. Let $k1$ and $k2$ be the indices representing the two new clusters created from splitting the kth cluster. Then, the split score L_k^{split} for each cluster is L_{k1} + L_{k2} - L_k where L_i defines the observation likelihood of the data in cluster i. The merge score for two template clusters $L_{k'k''}^{merge}$ is computed by defining the likelihood score based on the center of the new cluster (indexed by $k'k''$) inferred from all time series assigned to both clusters k' and k'' being merged. Thus, $L_{k'k''}^{merge} = L_{k'k''} - L_{k'} - L_{k''}$. A split-merge step with candidate clusters (k, k', k'') is accepted if $L_k^{split} + L_{k'k''}^{merge} > 0$.[5] Note that the merge and split steps can also be used independently to reduce or increase the number of CSTs in the model by trading it off against an appropriate penalty. In our implementation, we use the merge and split steps together to preserve the CST count and primarily use these steps as a way of escaping local maxima.

4.3.4 Model Initialization

A known problem with EM is its susceptibility to local-maxima and dependence on a good initialization [Koller and Friedman, 2009]. One approach to initializing the model is a *window-length based approach*. Prior works have used variants of this approach [Minnen *et al.*, 2007]. Given a pre-specified window length, using a sliding window, the signal can be windowed. Clusters extracted using k-means on these windowed signals can be used as an initialization for CSTM.

The choice of window length is not always obvious, especially in domains where motifs show considerable warp. An alternative approach is to describe the desired motifs in terms of their structural complexity — the number of distinct extrema points (peaks and dips) in the motif. For example, the signals in figure 4.7b, figure 4.9a and figure 4.9b have 4,

[5] In order to avoid curve fitting exhaustively to all candidate pairs for the merge move, we propose plausible pairs based on the distance between their template means (s^k), and then evaluate the benefit of the merge-split step using the correct objective. Candidate pairs with the smallest distance between their template means are considered first.

8 and 6 extrema points. Such an extrema profile is desirable because it provides a warp invariant signature of the motif. Simple k-means clustering based on these peak profiles yields initializations where clusters have segments with similar shapes yet varying lengths. We call this approach *peak-based initialization*. We describe this method in more detail below.

Peak-based initialization

We first characterize a segment s in the continuous time series by the set of its extremal points [Fink and Gandhi, 2010] — their heights $f_1^s, \cdots, f_M^s$ and positions $t_1^s, \ldots, t_M^s$.[6] We can find segments of the desired structure using a simple sliding window approach, in which we extract segments s that contain the given number of extremal points (we only consider windows in which the boundary points are extremal points). We now define $\delta_m^s = f_m^s - f_{m-1}^s$ (taking $f_0^s = 0$), that is, the height difference between two consecutive peaks. The *peak profile* $\delta_1^s, \ldots, \delta_M^s$ is a warp invariant signature for the window: two windows that have the same structure but undergo only temporal warp have the same peak profile. Multidimensional signals are handled by concatenating the peak profile of each dimension. We now define the distance between two segments with the same number of peaks as the weighted sum of the L_2 distance of their peak profile and the L_2 distance of the times at which the peaks occur:

$$d(s, s') = \sum_{m=1}^{M} \|\delta_m^s - \delta_m^{s'}\|_2 + \eta \sum_{m=1}^{M} \|t_m^s - t_m^{s'}\|_2 \tag{4.11}$$

The parameter η controls the extent to which temporal warp is considered in the similarity metric (for example, $\eta = 0$ defines an entirely warp-invariant distance); we use $\eta = 1$. Using the metric d, we cluster segments (using e.g., k-means) and select the top K most compact clusters as an initialization for CSTM. Compactness is evaluated as the distance between all segments in the cluster to a single segment in the cluster, minimized over the choice of this segment.

[6] Noisy peaks below a threshold variation are removed by the extremal point detection algorithm of Fink and Gandhi [2010].

```
1. Subsample data.
2. Initialize clusters for CSTM.
     // Peak-based initialization or,
     // Window-length based initialization
3. Repeat until maximum number of iterations is reached or change in
   likelihood is less than a tolerance,
     a. E-step using Viterbi,
          // Optional optimization: Prune to keep highest probability
          // states at each time step during the forward step.
     b. M-step to estimate model parameters,
     c. Every few iterations, call split and merge to escape local maxima.
          // Optional optimization: Consider candidate pairs based
          // on distances between template means as a fast heuristic.
```

Table 4.2: Summary of the learning algorithm for CSTM

4.3.5 Preprocessing

The complexity of our algorithm depends linearly on the length of the series. Subsampling at the Nyquist frequency [Nyquist, 1928] lowers the sampling rate without loss of information in the data.[7] We did not build an automated procedure for identifying the Nyquist rate, but instead examined the Fourier coefficients of the series to select a threshold frequency greater than which the coefficients had little power. For the NICU dataset used here, the subsampling provided a more than a $10X$ speedup.

4.3.6 Summary of learning algorithm

We briefly summarize in table 4.2 the steps of the learning algorithm for CSTM.

4.4 Experiments and Results

We now evaluate the performance of CSTM for uncovering repeated shapes and compare it to other motif discovery methods. We evaluated CSTM on four different datasets. We compare on both classification accuracy and coverage, comparing to the widely-used pipeline that uses random-projection based methods of Mueen et al. [2009]; Chiu et al. [2003]. We also compare against variants of our model to elucidate the importance of novel bias our

[7] Intuitively, the Nyquist frequency is the highest frequency at which there is still information in the signal.

model imposes over prior work. We give a brief overview of our experimental setup before describing our results.

4.4.1 Baseline Methods

There are two different approaches that have been used for solving the temporal motif discovery problem.

We refer to the baselines based on the first approach as **GreedyRPM** and **GreedyRPC**. These are motivated by random-projection based pipelines widely used in prior work. For GreedyRPM, first the series is windowed using a sliding window approach. A random projection based algorithm [Mueen *et al.*, 2009] repeatedly finds the *closest matched pair* of windows from the set of all candidate windows as a motif. To avoid finding similar motifs at consecutive iterations, we remove candidate windows within distance d times the distances between the closest pair at every iteration. Unlike the CSTM, which optimizes a global objective to discover the most frequent motifs, this pipeline greedily discovers the most closely matched pairs as motifs one at a time. Thus, we refer to this method as GreedyRPMd (for Greedy Random-Projection method using the motif discovery procedure of Mueen et al. [2009] and d specifies the degree of variability allowed within a cluster, larger d leading to larger variability). For GreedyRPC, the procedure in Mueen et al. [2009] is replaced with the procedure in Chiu et al. [2003]. The latter selects a motif at each iteration not merely based on closeness, but based on its frequency in the data. Clusters with a large number of assigned windows are selected first and a subsequence in the cluster within distance d of of the highest number of other subsequences in that cluster is selected as the motif. For both methods, to extract matches to existing motifs on a test sequence, at each point, we compute the distance to all motifs at all shifts and label the point with the closest matched motif.

A second series of approaches optimizes the model likelihood, albeit with different modeling biases. To understand the contribution to the performance for each of our model biases, we define variants of the CSTM where we use alternative biases that have been previously used. Minnen et al. [2007] and others have used dynamic time warping for computing similarity between warped subsequences; we define the variant **CSTM-DTW** where dynamic time warping is used instead of the structured warp matrices within CSTM. Each row of the warp matrix is fixed to be the uniform distribution.

A different variant **CSTM-NW** allows no warps. This is done by setting the warp

matrix to be of size $W = 1$. We also define the variant **CSTM-MC** which represents the motif as a simple template encoded as a mean vector (one for each point), as done in majority of prior works [Oates, 2002; Minnen *et al.*, 2007; Denton, 2005] instead of enforcing the smoothness bias on the motifs.

The goal of the peak-based initialization procedure is to provide an alternative to window-length methods in domains where the length of the motif is less obvious. Therefore, to evaluate the competitiveness we report results using this initialization procedure as **CSTM-PB**.

4.4.2 Metric

To evaluate the quality of the recovered templates, we use classification accuracy as our primary metric. We treat the discovered motifs as the feature basis and their relative proportions within a segment as the feature vector for that segment. Thus, for each true motif class (e.g., a character or action) a mean feature vector is computed from the training set. On the test set, each true motif is assigned a label based on distance between its feature vector and the mean feature vector for each class. Classification performance on the test set are reported. This way of measuring accuracy is less sensitive to the number of templates used.[8]

4.4.3 Datasets

We use four datasets for validation of our model. Although our primary motivation for this work arose from clinical data, it is difficult to establish ground truth in clinical data. As discussed in chapter 1, labeling these datasets requires valuable clinician time, which makes the expense of labeling this data prohibitive. Moreover, tasks of discovery rather than recognition are more valuable in this domain as the definition of ground truth is not always well-established. We select two other real-world multivariate datasets where the motifs are labeled upfront.

The *Character* data is a collection of x and y-pen tip velocities generated by writing

[8]To the reader who is curious about why the regime used for measuring phoneme or word segmentation performance in the speech literature could not be used directly, classification is used as the metric of choice there. However, typically labeled segments are available to train a classification model and therefore, features are extracted from pre-segmented ground truth data. In our setting, we wish to evaluate the quality of segmentation generated on the training set. Thus, we construct the classification model on the segmentations extracted from the CSTM model in an unsupervised manner.

characters on a tablet [Keogh and Folias, 2002]. We concatenated the individual series to form a set of labeled unsegmented data for motif discovery.[9]

The *Kinect Exercise* data was created using Microsoft Kinect^{TM} by tracking a subject doing leg-exercises. The data features six leg exercises such as front kick, rotation, and knee high, interspersed with other miscellaneous activity as the subject relaxes. There are six to fifteen repetitions of each activity; these repeated instances do not appear all at once. The dataset was collected in two different settings. We extract the three dimensional coordinates of the ankle. Different exercises appear as motifs in this three-dimensional stream.

A third *Simulated* dataset of seven hand-drawn curves was used to evaluate how model performance degrades under different amounts of NRW segments. In figure 4.2, we show the hand-drawn curves. With these templates and a random intialization of our model, we generated four different datasets where the proportions of non-repeating segments were 10%, 25%, 50% and 80%.

The final dataset we use is a clinical dataset from University of Edinburgh. The dataset has heart rate data collected from 15 infants in the Neonatal Intensive Care Unit (*NICU*) [Williams *et al.*, 2005b]. The data for each infant was collected every second for 24 hours, on nine channels: heart rate, systolic and diastolic blood pressures, $TcPO_2$, $TcPCO_2$, O_2 saturation, core temperature and incubator temperature and humidity. Our goal here is to discover temporal events that are indicative of health status of the neonate. We chose this dataset because it has been labeled by a neonatologist; only one clinical event called bradycardia is marked. Other labels included events such as "incubator open" or "temperature probe disconnection" which, although useful for reducing false alarms, are not relevant to our goal of linking events to health status so we ignore these events. The dataset was labeled by the neonatologist with the intent to demonstrate many variations of any given event in the data, as opposed to comprehensively labeling every instance of that event. Moreover, there may be repeating events in the data not known to the neonatologist and hence, not marked. Therefore, this dataset, unlike the previous three, is partially labeled. For this reason, we focus on measuring specificity and sensitivity of recovering the known labelled events of bradycardia as well as uncovering any novel shapes.

[9]The dataset also has pen-tip pressures. The pen-tip pressure has obvious structure which is that the pressure goes to zero at the end of each characrer. This makes the segmentation problem easy for a model based method like CSTM so we do not include the pressure channel in our dataset.

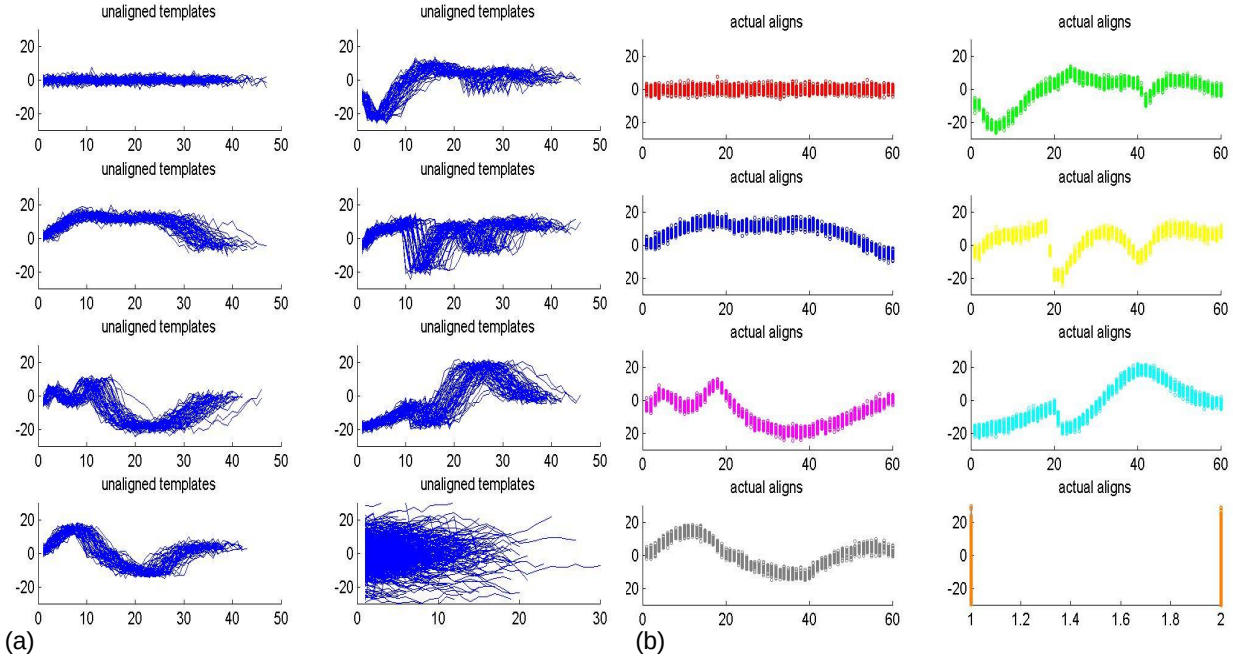


Figure 4.2: Hand-drawn curves from the simulated dataset. a) Example sequences generated from each of the motifs, b) Examples sequences aligned to their canonical shape template. The bottom right subplot in (a) shows that sequences from the NRW state do not have the same repeating structure as those from the CSTs do. The bottom right suplot from (b) shows the data once aligned to the NRW state. Each segment is of unit length once aligned.

4.4.4 Results

Our method, similar to prior work, requires an input of the template length and the number of templates K. When characterizing the motif length is unintuitive, peak based initialization can be used to define the initial templates based on complexity of the desired motifs. In addition, our method requires a setting of the NRW self-transition parameter: λ controls the tightness of the recovered templates and can be incrementally increased (or decreased) to tune to desiderata.

The intuition for how λ must be set can be derived by considering relevant terms in the likelihood. A data point y_t is assigned to a template versus the NRW state based on whether the likelihood of assignment to the former is larger than the latter. More specifically,

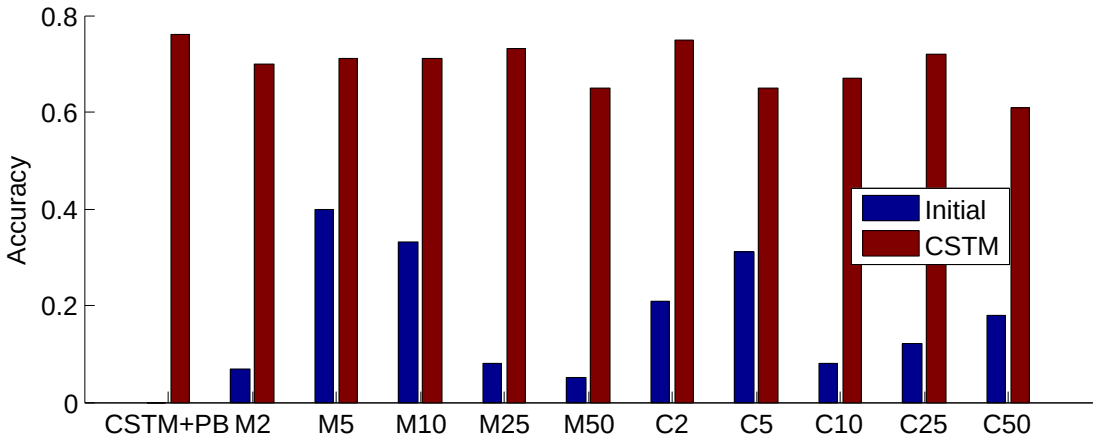


Figure 4.3: Comparison of the CSTM with an initialization using the peak-based method, and initializations from GreedyRPM and GreedyRPC with different settings for d on the character dataset. In the figure, GreedyRPM and GreedyRPC have been abbreviated as Md and Cd respectively.

$$\log \pi_{\omega}^{\kappa_t} - 0.5 \log 2\pi\dot{\sigma}^2 - \frac{(y_t - \phi_t s^{\kappa_t}(\rho_t))^2}{\dot{\sigma}^2} > \log \lambda - 0.5 \log 2\pi\sigma^2 - \frac{(y_t - y_{t-1})^2}{\sigma^2}$$

Assuming the data is explained just as well by the template and the NRW state, we get:

$$\frac{(y_t - \phi_t s^{\kappa_t}(\rho_t))^2}{\dot{\sigma}^2} - \frac{(y_t - y_{t-1})^2}{\sigma^2} \approx 0$$

Therefore,

$$\log \lambda < \log \pi_{\omega}^{\kappa_t} - 0.5 \log \frac{\dot{\sigma}^2}{\sigma^2}$$

$$\lambda < \pi_{\omega}^{\kappa_t} \frac{\sigma}{\dot{\sigma}}$$

Say π_{ω} is Uniform$(1/W, ..1/W)$ where W is the width of the warp matrix. Then,

$$\lambda < (1/W) * \frac{\sigma}{\dot{\sigma}}$$

In all our experiments, we set $\lambda = 0.5$, a value which respects this constraint. For the choice of configurations for the various transformation models, CSTM only exploits the temporal warp and local shift transformations in the experiments below. Our data showed little amplitude variation so we did not experiment with the amplitude transformation model. Three different warp levels were allowed i.e. $W = 3$.

Character Data. On the character dataset, for different settings of the parameters, number of clusters and the distance d, we computed classification accuracies for GreedyRPM and GreedyRPC. The window length is easy to infer for this data even without knowledge of the actual labels; we set it to be 15 (in the subsampled version). We experiment with different initializations for CSTM: using the motifs derived by the methods of GreedyRPM and GreedyRPC, and those derived using the peak-based initialization. Figure 4.3 shows the classification accuracies for these different initializations. The performance of a random classifier for this dataset is 4.8%. Our method consistently dominates GreedyRPM and GreedyRPC by a large amount and yields average and best case performance of 68.75% and 76.25% over all initializations. Our method is also relatively insensitive to the choice of initialization. Our best performance is achieved by initializing with the peak-based method (**CSTM+PB**) which requires no knowledge of the length of the template. Moreover, for those parameter settings where GreedyRPM does relatively well, our model achieves significant gain by fitting warped versions of a motif to the same template. In contrast, GreedyRPM and GreedyRPC must expend additional templates for each warped version of a pattern, fragmenting the true data clusters and filling some templates with redundant information, thereby preventing other character patterns from being learned. Increasing the distance parameter for GreedyRPM can capture more warped characters within the same template; however, many characters in this dataset are remarkably similar and performance suffers from their misclassification as d increases.

In the next batch of experiments, we focused on a single initialization. Since GreedyRPM performed better than GreedyRPC, and is relatively more stable, we consider the GreedyRPM10 initialization, and compare CSTM against its variants with no warp, with uniform warp, and without a template prior. In figure 4.4a, we see that performance degrades in all cases. A qualitative examination of the results showed that, while the no-warp version fails to align warped motifs, the DTW model aligns too freely resulting in convergence to poor optimum. This becomes evident when the confusion matrices for two models are compared (as shown in figure 4.5). We see that the CSTM-DTW has many more off-diagonal terms

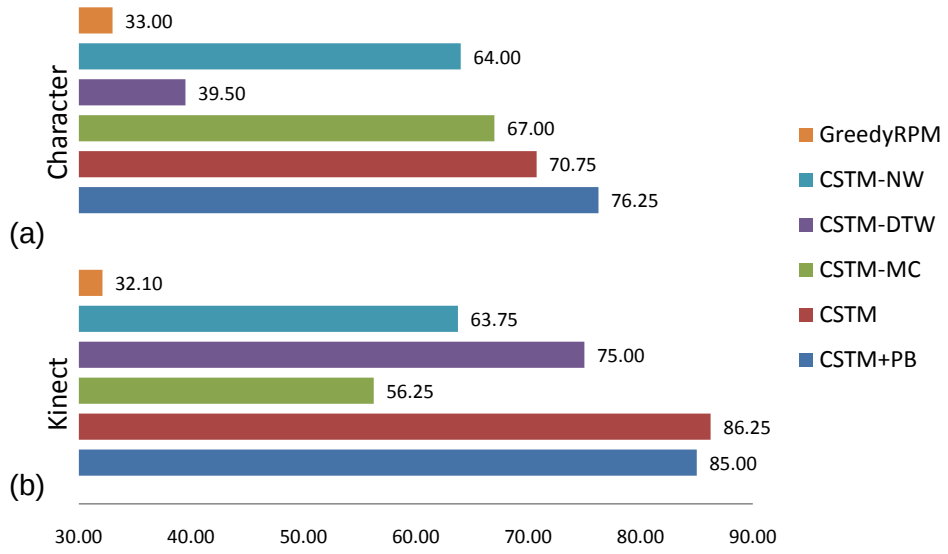


Figure 4.4: Accuracy on Character (top) and Kinect (bottom) for CSTM and its variants. Two different initializations for CSTM are compared: GreedyRPM10 and peak-based.

that are positive reflecting confusion between the corresponding characters are confused by CSTM-DTW; in comparison, CSTM (with a structured warp matrix) has far fewer off-diagonal terms. Where CSTM misses, we see that it fails in intuitive ways, with many of the misclassifications occurring between similar looking letters, or letters that have similar parts; for example, we see that h is confused with m and n, p with r and w with v.

Kinect Data. Next, we tested the performance of our method on the Kinect Exercise dataset. To evaluate GreedyRPM on this dataset, we tested GreedyRPM with parameter settings taken from the cross product of template lengths of 5, 10, 15, or 20, distance thresholds of 2, 5, 10, 25, or 50, and a number of templates of 5, 10, 15, or 20. A mean accuracy of 20% was achieved over these 80 different parameter settings; accuracies over 50% were achieved only on 7 of the 80, and the best accuracy was 62%. Using GreedyRPM10 as an initialization (with 10 clusters and window length 10, as above), we evaluate CSTM and its variants. CSTM achieves performance of over 86%, compared to the 32% achieved by GreedyRPM10 directly. CSTM with a peak-based initialization (using either 5 or 7 peaks) produced very similar results, showing again the relative robustness of CSTMs to

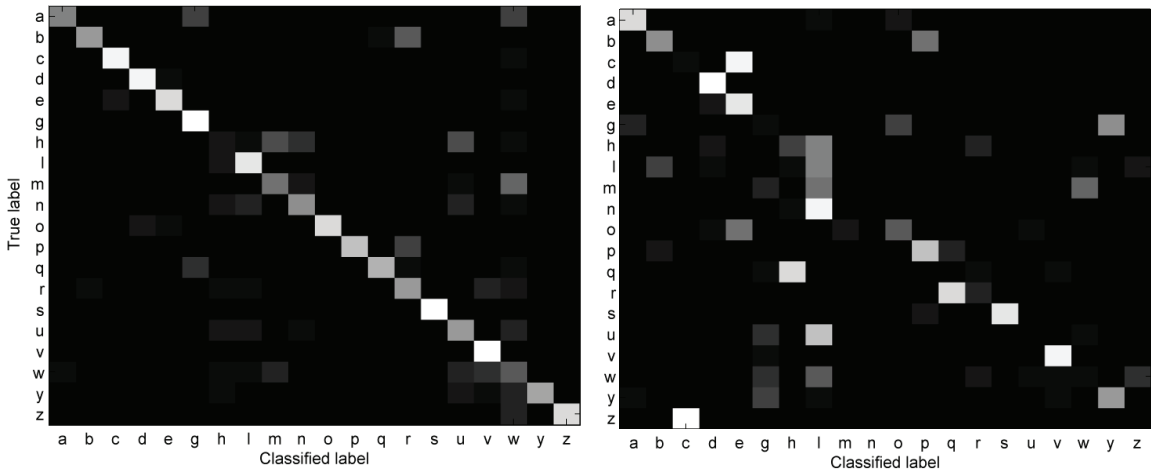


Figure 4.5: Confusion matrix showing performance of CSTM (left) and CSTM-DTW (right) on the character data.

initialization. Comparing to different variants of CSTM, we see that the lack of bias in the template representation in this dataset lowers performance dramatically to 56.25%. We note that the templates here are relatively short, so, unlike the character data, the drop in performance due to unstructured warp is relatively smaller.

Synthetic Data. To evaluate how our model performs as the proportion of non-repeating segments increases, we evaluate the different variants of CSTM and GreedyRPM10 on simulated data of hand-drawn curves. CSTM performance is 78% even at the 80% random walk level, and performs considerably better than GreedyRPM10, whose performance is around 50% (see figure 4.6). Although the performance gap between CSTM and its less-structured variants is largest at the lower NRW levels, CSTM's performance ($74.5\% - 90\%$) remains consistently higher than these other models ($62\% - 72\%$).

NICU Data. For the NICU data, in figure 4.7a and figure 4.7b, we show example clusters containing bradycardia signals generated by GreedyRPM and CSTM respectively. The former is able to capture highly variable versions of bradycardia while those in the latter are fairly homogeneous. In figure 4.7c, we show aligned versions of the signals in figure 4.7b; once aligned, its more visually evident that the CSTM cluster sequences are true bradycardia subseqeunces.

In figure 4.8, we show the ROC curve for identifying bradycardia using each of the

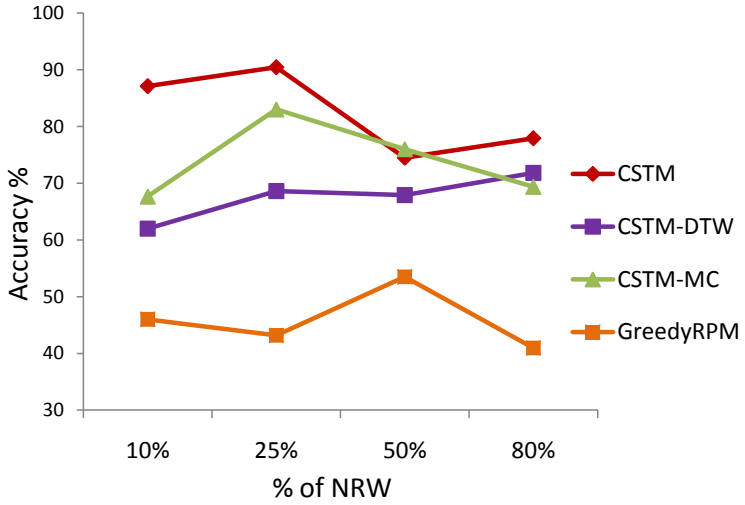
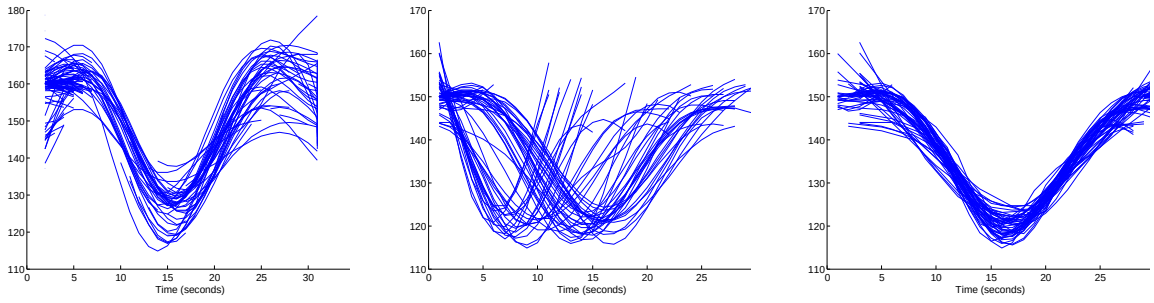


Figure 4.6: Classification performance for increasing level's of NRW to CST proportion in the data.



a. b. c.

Figure 4.7: a) An example bradycardia cluster extracted by GreedyRPM, b) an example bradycardia cluster recovered by CSTM, c) the bradycardia cluster shown with sequences aligned to the shape template. Note that the bradycardia sequences extracted by CSTM are more heterogeneous in appearance than those captured by GreedyRPM. Thus, CSTM is able to better capture the variability in the cluster.

models. True positive and false positive measures[10] are computed as each new cluster is added up to a total of 20 clusters. For each ROC computation, clusters are added in order

[10]True positive and false positive are measured per data point.

of their individual f1 scores.[11] We perform a single run with peak based clustering using $3 - 7$ peaks and multiple runs for GreedyRPM with different settings for the parameter d (see figure 4.8). The ROC curve from CSTM dominates those from GreedyRPM with significantly higher true positive rates at lower false positive rates. In figure 4.9a and figure 4.9b, we show examples of novel clusters not previously known (and potentially clinically significant).[12]

4.5 Discussion and Conclusion

In this chapter, we have presented a new model for unsupervised discovery of deformable motifs in continuous time series data. Our probabilistic model seeks to explain the entire series and identify repeating and non-repeating segments. This approach allows us to model and learn important representational biases regarding the nature of deformable motifs. We demonstrate the importance of these design choices on multiple real-world domains, and show that our approach performs consistently better compared to prior works.

Our work can be extended in several ways. Our warp-invariant signatures can be used for a forward lookup within beam pruning to significantly speed up inference when K, the number of templates is large. Our current implementation requires fixing the number of clusters. However, our approach can easily be adapted to incremental data exploration, where additional templates can be introduced at a given iteration with split and merge to refine existing templates or discover new templates. A Bayesian nonparametric prior is another approach (e.g., [Richardson and Green, 1997]) that could be used to systematically control the number of classes based on model complexity. A different extension could build a hierarchy of motifs, where larger motifs are comprised of multiple occurrences of smaller motifs, thereby possibly providing an understanding of the data at different time scales. More broadly, this work can serve as a basis for building non-parametric priors over deformable multivariate curves.

[11] Under this metric, large pure clusters increase performance on AUC more than smaller pure clusters, which is desirable. Moreover, as expected, pure bradycardia clusters increase AUC more than clusters with a large number of false positives.

[12] When these clusters were shown to a neonatologist expert at Stanford, she remarked that the cluster in figure 4.9b is a significant event and a respiratory intervention followed by chest compressions should have occurred.

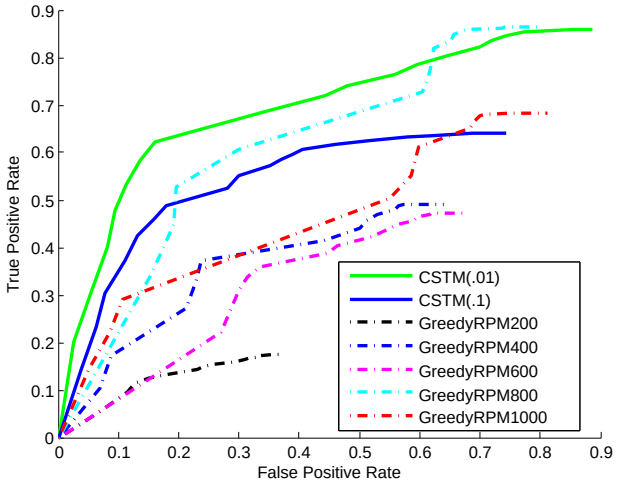
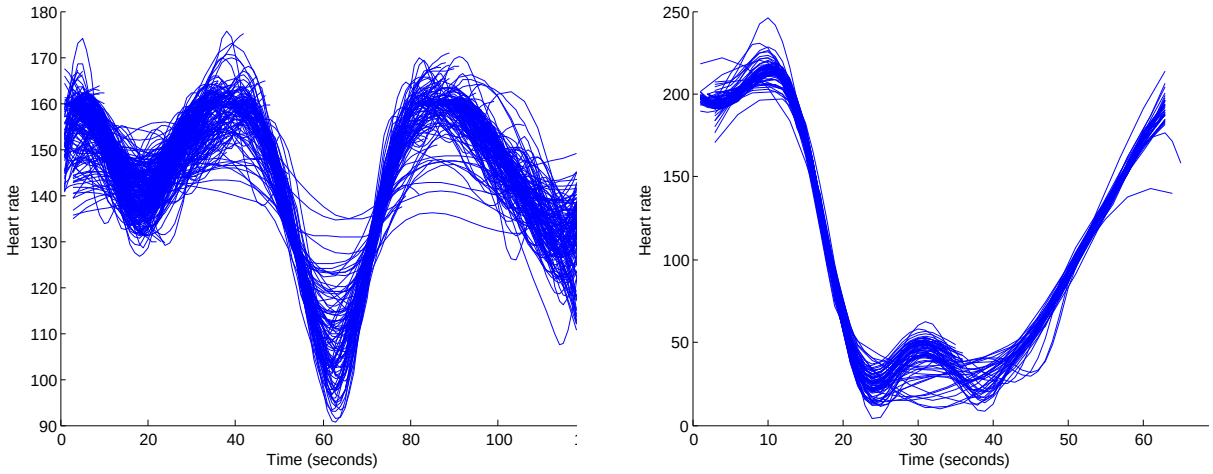


Figure 4.8: ROC curve for recovering bradycardia sequences from the data using CSTM and GreedyRPM with various parameter settings for both models.



a. b.

Figure 4.9: Two examples of novel clinical events, a) mutliple varying level's of bradycardia like episodes (bradycardia in a preemie is defined as HR < 80 for longer than 10 seconds) within a short span of 120 seconds, b) fast and large oscillations in heart rate.

Chapter 5

Clinical Application to Risk Stratification

Using cues for infant health extracted from the physiologic signals, we aimed to develop a modern tool akin to an electronic Apgar assessment that reflects a newborns physiological status and is predictive of future illness severity. Physiological data are routinely recorded in intensive care, but their use for rapid assessment of illness severity or long-term morbidity prediction has been limited. In this chapter, we describe the development of a physiological assessment score for preterm newborns based on standard signals recorded noninvasively on admission to a neonatal intensive care unit. We were able to accurately and reliably estimate the probability of an individual preterm infants risk of severe morbidity on the basis of noninvasive measurements alone. This prediction algorithm was developed with electronically captured physiological time series data from the first 3 hours of life in preterm infants (34 weeks gestation, birth weight $\leq$ 2000 g). Extraction and integration of the data with state-of-the-art machine learning methods produced a probability score for illness severity, the PhysiScore. We validated PhysiScore on 138 infants with the leave-one-out method to prospectively identify infants at risk of short- and long-term morbidity. PhysiScore provided higher accuracy prediction of overall morbidity (86% sensitive at 96% specificity) than other neonatal scoring systems, including the standard Apgar score. PhysiScore was particularly accurate at identifying infants with high morbidity in specific complications (infection: 90% at 100%; cardiopulmonary: 96% at 100%). Physiological parameters, particularly short-term variability in respiratory and heart rates, contributed more to morbidity prediction than invasive laboratory studies. Our flexible methodology of individual risk prediction

based on automated, rapid, noninvasive measurements can be easily applied to a range of prediction tasks to improve patient care and resource allocation.

5.1 Introduction

Early, accurate prediction of a neonate's morbidity risk is of significant clinical value because it allows for customized medical management. The standard Apgar score has been used for more than 50 years to assess neonatal well-being and the need for further medical management. Based on quantitative and qualitative characteristics of appearance, pulse, grimace, activity and respiration, the Apgar is computed as a score between 1 and 10, 10 signifying the infant is in good health. A low score is indicative of poor health and his need for medical attention. We aimed to develop a modern tool akin to an electronic Apgar assessment that is fully non-invasive, reflects a newborns physiological status and is predictive of future illness severity. Such an improvement in neonatal risk stratification may better inform decisions regarding aggressive use of intensive care, need for transport to tertiary centers, and resource allocation, thus potentially reducing the estimated $26 billion per year in U.S. health care costs resulting from preterm birth [Behrman and Butler, 2007].

Previously, various risk factors have been associated with assessing neonatal health. Gestational age and birth weight are highly predictive of death or disability [Robertson *et al.*, 1992] but do not estimate individual illness severity or morbidity risk [Tyson *et al.*, 2008]. These perinatal risk factors, in addition to laboratory measurements, have been incorporated into currently used algorithms for mortality risk assessment of preterm infants [Richardson *et al.*, 1993; 2001; Network, 1993]. These algorithms, however, predict mortality rather than morbidity [Tyson *et al.*, 2008]. They also rely on invasive testing and require extraction of data from multiple sources to make a risk assessment.

Although it has been recognized that changes in heart rate characteristics [Schulte-Frohlinde *et al.*, 2002] or variability [Tsuji *et al.*, 1994] can suggest impending illness and death in a range of clinical scenarios, from sepsis [Griffin *et al.*, 2005a; 2005b] in intensive care patients to fetal intolerance of labor [Williams and Galerneau, 2003a], the predictive accuracy of a single parameter is limited. Intensive care providers observe multiple physiological signals in real time to assess health, but certain informative patterns may be subtle.

To achieve improved accuracy and speed of individual morbidity prediction for preterm neonates, we developed a new probability score (PhysiScore) based on physiological data

obtained noninvasively after birth along with gestational age and birth weight. We evaluated PhysiScores use for predicting overall morbidity and mortality, specific risk for infants with infection or cardiovascular and pulmonary complications, and a combination of complications associated with poor long-term neurodevelopment and compared its performance to standard scoring systems in a preterm neonatal cohort.

5.2 Risk Stratification

The goal of risk stratification is to quantify an infant's overall risk for major complications and identify which infants will require intensive clinical attention. For this purpose, patients in our study are classified into two groups – high morbidity (HM) or low morbidity (LM) – on the basis of their illnesses.

The HM group was defined as any patient with major complications associated with short- or long term morbidity. These are patients that typically need more intensive support in the NICU. More specifically, short-term morbidity complications included culture-positive sepsis, pulmonary hemorrhage, pulmonary hypertension, and acute hemodynamic instability. Long-term morbidity was defined by moderate or severe bronchopulmonary dysplasia (BPD), retinopathy of prematurity (ROP) stage 2 or greater, intraventricular hemorrhage (IVH) grade 3 or 4, and necrotizing enterocolitis (NEC) on the basis of the strong association of these complications with adverse neurodevelopmental outcome. Death was also included in the long-term morbidity group. Most infants in the HM category had short- and long-term complications affecting multiple organ systems.

Infants with none or only common problems of prematurity such as mild respiratory distress syndrome (RDS) and patent ductus arteriosus (PDA) without major complications were classified as LM.

5.3 Methods

Now, we describe each of the building blocks used for arriving at our non-invasive risk stratification score. We built a framework that (i) processes physiological parameters using nonlinear models, (ii) uses regularization to do automatic feature selection, and (iii) combines relevant features using multivariate logistic regression to produce the predictive score.

5.3.1 Feature construction

Our choice of input sources to be considered were derived based on three different sources of motivation. First, an early iteration of the TSTM on a small subset of patients (n=12) yielded insight that simple markers (such as those in discussed in Chapter 3) computed from physiologic signals are indicative of downstream morbitidy. Second, based on laboratory measurements that had been incorporated in previous risk stratification studies [Richardson *et al.*, 2001; Network, 1993], we extracted measurements of white blood cell count, band neutrophils, hematocrit, platelet count, and initial blood gasmeasurement of PaO_2 (partial pressure of oxygen, arterial), $PaCO_2$ (partial pressure of carbon dioxide, arterial), and pH(if available at < 3 hours of age). Third, based on a priori clinical knowledge, features of birth-weight, gestational age and signal means are extracted.

5.3.2 Physiologic signal processing

Time series heart rate, respiratory rate, and oxygen saturation data are collected from all CR monitors. The data for this study was recorded at the minute granularity. Heart and respiratory rate signals are processed to compute a base and residual signal. The base signal represents a smoothed, long-term trend; it is computed with a moving average window of 10 minutes. The residual signal is obtained by taking the difference between the original signal and the base signal; it characterizes short-term variability most likely linked to sympathetic function (see Figure 5.1 for an illustration).

Physiological signals recorded in the first 3 hours of life. This time period was selected for analysis because it is less likely to be confounded by medical interventions and provides prediction early enough in the infant's life to be useful for planning therapeutic strategy. For heart and respiratory rates, we compute the base signal mean, base signal variance, and residual signal variance. The variance features were motivated by analysis from Chapter 3 on our preliminary set of 12 patients. For the oxygen saturation, we compute the mean and the ratio of hypoxia (oxygen saturation $< 85\%$) to normoxia.

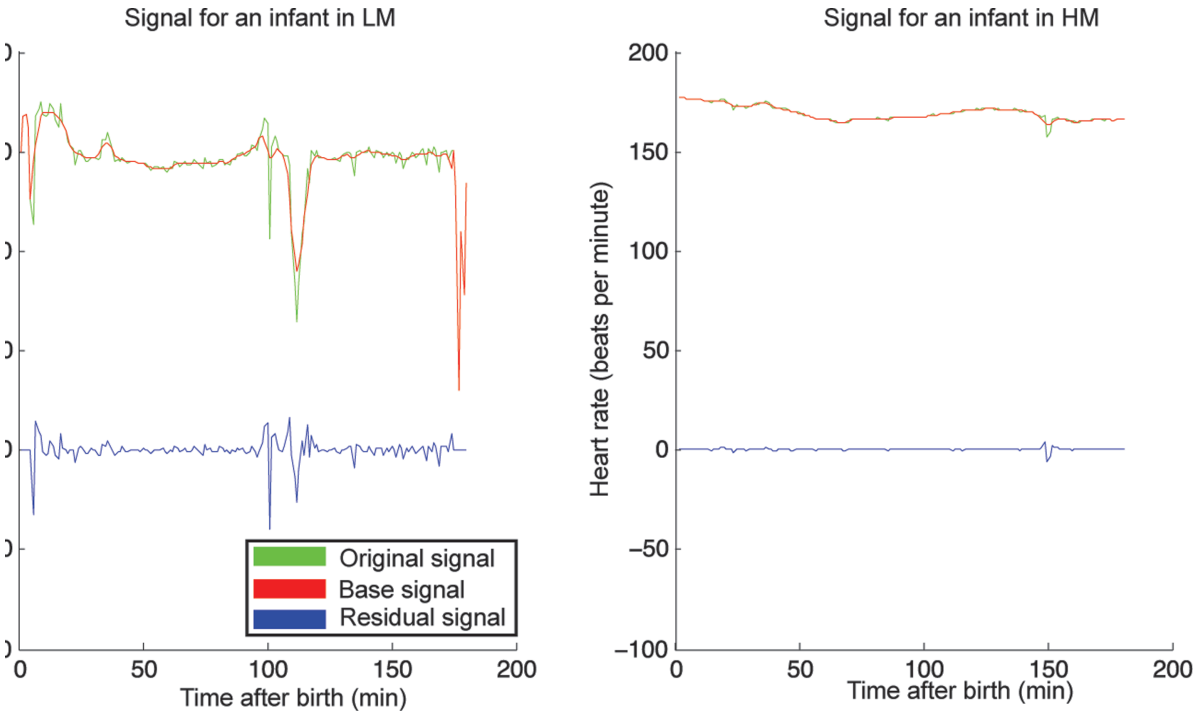


Figure 5.1: Processing signal subcomponents. Differing heart rate variability in two neonates matched for gestational age (29 weeks) and weight (1.15 ± 0.5 kg). Original and base signals are used to compute the residual signal. Differences in variability can be appreciated between the neonate predicted to have HM (right) versus LM (left) by PhysiScore.

5.3.3 Combining risk factors into a score

Individual risk factors are aggregated using a logistic regression model as

$$\mathrm{P(HM}|v_1, v_2, \cdots, v_n) = (1 + \exp(b - w_0 c - \sum_{i=1}^{n} w_i f(v_i)))^{-1} \tag{5.1}$$

where n is the number of risk factors and c = log P(HM)/P(LM) is the a priori log odds ratio of the classes high morbidity (HM) and low morbidity (LM). The ith characteristic, v_i (e.g., physiological parameter or weight) is used to derive a numerical risk feature $f(v_i)$ via nonlinear Bayesian modeling (which we discuss in detail in section 5.3.5).

The score parameters b and $\vec{w}$ are learned from the training data for use in prospective risk prediction. The parameter w_i represents the weight of the contribution of the ith characteristic to the computed probability score, with higher weight characteristics having a greater effect.

5.3.4 Learning the score parameters

To learn the score parameters b and $\vec{w}$ where there are a total of m features, we maximize the log likelihood of the data in the training set with a ridge penalty as

$$\underset{w,b}{\operatorname{argmax}} \Sigma_{j=1}^{n} \log P(\mathrm{HM}|v_1^j, \cdots, v_m^j) - \lambda \Sigma_{i=0}^{m} w_i^2 \tag{5.2}$$

The ridge penalty reduces spurious data dependence and prevents overfitting by controlling model parsimony [Hastie *et al.*, 2001; Zhu and Hastie, 2004]. The hyperparameter λ controls the complexity of the selected model.

5.3.5 Nonlinear models of risk factors

To implement Eq. 5.1, we must determine how to integrate the various risk factors — continuous and discrete — including the physiological measurements, into our risk model. Several approaches exist in the literature. One common approach is to define a normal range for a measurement and use a binary indicator whenever the measurement is outside that range. Although this approach can most easily be implemented in a clinical setting, it provides only coarse-grained distinctions derived from extreme values. Another approach is to predetermine a particular representation of the continuous valued measurement, usually either the feature itself, or a quadratic or logarithmic transformation, as selected by an expert [Whitlock *et al.*, 2009; Schnabel *et al.*, 2009].

We use a different approach based on a Bayesian modeling paradigm [Ross, 2004]. This approach captures the nonlinear relationships between the risk factor and the outcome and takes into account that the overall behavior of a factor can vary greatly between sickness categories. For each risk factor v_i, we separately learn a parametric model of the distribution of observed values in the training set $P(v_i|C)$ for each class of patient C (HM and LM). The parametric model is selected and learned with maximum-likelihood estimation (Figure 5.2) from the set of long-tailed probability distributions of exponential, Weibull, lognormal, normal, and gamma. Specifically, for each parametric class, we fit the maximum likelihood

parameters and then select the parametric class that provides the best (highest likelihood) fit to the data. The log odds ratio of the risk imposed by each factor is incorporated into the model as denoted by $f(v_i)$ in Eq. 5.1.

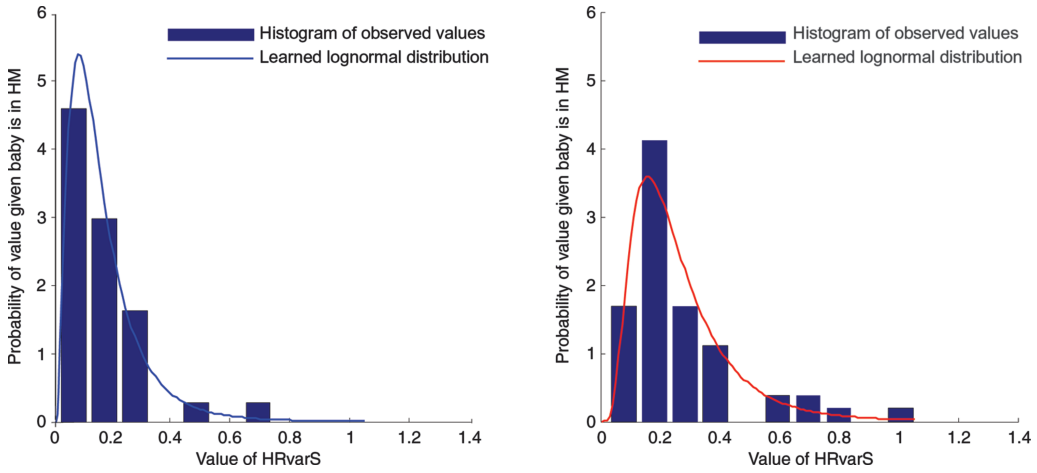


Figure 5.2: Distribution of residual heart rate variability (HRvarS) in all infants. Learned parametric distributions overlaid on the data distributions for HRvarS displayed for the HM versus LM categorization.

An important advantage of our approach is that explicit missing data assumptions can be incorporated. When standard laboratory results (for example, complete blood count) are not recorded, we assume that they are missing at random and not correlated with outcome. Their contribution if missing is 0 and $\log P(v_i|\text{HM})/P(v_i|\text{LM})$ otherwise. Blood gas measurements, however, are likely obtained only for profoundly ill patients and hence are not missing at random. Thus, for each measurement type i, we define $m_i = 1$ if measurement v_i is missing and $m_i = 0$ otherwise. We now learn the distribution $P(m_i|C)$ (the chance that the measurement i is missing for each patient category C) and $P(v_i|C, m_i = 0)$ (the distribution of the observed measurements) as described above. The factor contribution for measurement i is computed as

$$f(v_i) = \begin{cases} \log \dfrac{P(v_i|\text{HM}, m_i = 0)}{P(v_i|\text{LM}, m_i = 0} + \log \dfrac{P(m_i = 0|\text{HM})}{P(m_i = 0|\text{LM})}, & \text{if } m_i = 0 \quad (5.3) \\ \log \dfrac{P(m_i = 1|\text{HM})}{P(m_i = 1|\text{LM})}, & \text{if } \ m_i = 1 \quad (5.4) \end{cases}$$

This formulation allows us to account both for the observed measurement, if present, and for the likelihood that a particular measurement might be taken for patients in different categories.

This approach has additional advantages. Putting all factors in a probabilistic framework provides a comparable representation for different risk factors, allowing them to be placed within a single, integrated model. Using a parametric representation of each continuous measurement alleviates issues arising from data scarcity. Uncovering the dependence between the risk factor and the illness category automatically reduces data requirement by eliminating the need for cross-validation to select the appropriate form. Unlike most previous methods, we used different parametric representations for patients in different categories, better capturing disease-induced changes in patient physiology. Finally, we obtained an interpretable visual summary of the likelihood of low patient morbidity over the range of values for each factor (shown in Figure 5.5B).

5.3.6 PhysiScore: Probabilistic score for illness severity

PhysiScore is a probability score that ranges from 0 to 1, with higher scores indicating higher morbidity. PhysiScore is calculated by integrating the following 10 patient characteristics into Eq. 5.1: mean heart rate, base and residual variability; mean respiratory rate, base and residual variability; mean oxygen saturation and cumulative hypoxia time; gestational age and birth weight. Each of these patient characteristics is modeled as described above and carries a specific learned weight, as denoted by w in Eq. 5.1.

5.4 Experiments and Results

We now evaluate the performance of PhysiScore. A gold-standard dataset with annotations validated by expert neonatologists was curated. We compare performance of PhysiScore with other state-of-the-art risk stratification scores. We give a brief overview of our experimental setup before describing our results.

5.4.1 Outcome annotation

Electronic medical records, imaging studies, and laboratory values were reviewed by pediatric nurses. In addition, annotations were verified in detail by an expert neonatologist. A second neonatologist was consulted when any ambiguities arose. All significant illnesses during the hospitalization were recorded. Morbidities were identified with previously described criteria: BPD [Ehrenkranz *et al.*, 2005], ROP [for the Classification of Retinopathy of Prematurity, 2005], NEC [Kliegman and Walsh, 1987], and IVH [Papile *et al.*, 1978]. For IVH and ROP, the highest unilateral grade or stage was recorded, respectively. Acute hemodynamic instability was also noted: hypotension (defined as a mean arterial blood pressure less than gestational age or poor perfusion) requiring ≤ 3 days of pressor support or adrenal insufficiency requiring hydrocortisone.

5.4.2 Study population

For this study, we used data from the inborn infants admitted to the NICU of Lucile Packard Children's Hospital from March 2008 to March 2009 who were eligible for enrollment. A total of 145 preterm infants met inclusion criteria: gestational age ≤ 34 completed weeks, birth weight ≤ 2000 g, and availability of cardiorespiratory (CR) monitor data within the first 3 hours of birth. Seven infants found to have major malformations were subsequently excluded.

Thus, to develop our prediction tool, we studied a total of 138 preterm neonates that were 34 weeks gestational age or less and < 2000 g in weight without major congenital malformations and with baseline characteristics and morbidities as shown in Table 5.1. Mean birth weight was 1367 g at an estimated mean gestational age of 29.8 weeks, placing these infants at significant risk of both short- and long-term complications.

Thirty-five neonates had HM complications. Of these, 32 had long term morbidities (moderate or severe BPD, ROP stage 2 or greater, grade 3 or 4 IVH, and/or NEC). Four neonates died after the first 24 hours of life. There were 103 preterm neonates with only common problems of prematurity (RDS and/or PDA) and so were considered LM. Five infants with a <2-day history of mechanical ventilation for RDS, but no other early complications, were transferred before ROP evaluation and marked as LM. Table 5.1 describes our infant population in detail.

5.4.3 Statistical methods

Sensitivity, specificity, AUC, and significance values [DeLong *et al.*, 1988] were computed for all comparisons. We used the leave-one-out method for all evaluations. With this method, predictive accuracy was evaluated for each patient separately. For each patient, we learned the model parameters with the data from all other patients as the training set and evaluated predictive accuracy on the held-out patient. This technique was repeated for each subject, as though each subject's clinical data was prospectively obtained. This method of performance evaluation is computationally intensive but is a well-established statistical method for measuring performance when the sample set size is limited [Rangayyan, 2005].

5.4.4 Results

Plotting the receiver operating characteristic (ROC) curve (Figure 5.3A) and associated area under the curve (AUC) values (Table 5.2) shows that PhysiScore exhibits good discriminative ability for prediction of morbidity and mortality risk and compares it to other risk assessment tools. Specifically, PhysiScore was compared to the Apgar score, long used as an indicator for the base physiological state of the newborn [Casey *et al.*, 2001], as well as to extensively validated neonatal scoring systems that require invasive laboratory measurements (Score for Neonatal Acute Physiology-II (SNAP-II) [Richardson *et al.*, 2001], SNAP Perinatal Extension-II (SNAPPE-II) [Richardson *et al.*, 2001], and Clinical Risk Index for Babies (CRIB) [Network, 1993]). For making predictions with the Apgar score, we constructed a model as in Eq. 5.1 using the 1- and 5-min Apgar scores as the only two inputs; this combined model outperformed either of the two Apgar scores when used in isolation. PhysiScore (AUC 0.9197) performed well across the entire range of the ROC curve and significantly better ($P < 0.003$) [DeLong *et al.*, 1988] than all four of the other comparison scores (Table 2). PhysiScore's largest performance gain occurred in the high-sensitivity/specificity region of the ROC curve (see region highlighted in Figure 5.3A (inset)). Setting a user-defined threshold based on desired sensitivity and specificity allows optimization for individual settings. For example, in our neonatal intensive care unit (NICU), a threshold of 0.5 achieves a sensitivity of 86% at a specificity of 95% for HM, as seen in Figure 5.3A (inset). Alternately, the use of a lower threshold would improve sensitivity at the expense of specificity.

We added the values obtained from laboratory tests to determine the magnitude of their

contribution to risk prediction beyond the PhysiScore alone (Figure 5.3B), incorporating parameters included in standard risk prediction scores (for example, SNAPPE-II): white blood cell count, band neutrophils, hematocrit, platelet count, and initial blood gasmeasurement of PaO2 (partial pressure of oxygen, arterial), $PaCO_2$ (partial pressure of carbon dioxide, arterial), and pH(if available at < 3 hours of age). No additional discriminatory power was achieved, suggesting that laboratory information is largely redundant with the patient's physiological characteristics.

To further assess the performance of PhysiScore, we analyzed prediction performance for infants in major morbidity categories. Specifically, we extracted two categories: infection (NEC, culture-positive sepsis, urinary tract infection, and pneumonia) (Figure 5.3C) and cardiopulmonary complications (BPD, hemodynamic instability, pulmonary hypertension, and pulmonary hemorrhage) (Figure 5.3D). Plotting data from the infants in the HM category who had a specific complication against data from all infants in the LM category yields ROC curves for discriminative ability for HM infants in these independent morbidity categories (Figure 5.3 C and D). Comparison to SNAPPE-II (the best-performing standard score) is also shown; AUCs were calculated for all scoring methods (Table 2) in these specifically defined sets. At a threshold of 0.5, PhysiScore achieves near-perfect performance (infection: 90% sensitivity at 100% specificity; cardiopulmonary: 96% at 100%).

Morbidity is most difficult to predict in patients with isolated IVH, for which all scores exhibit decreased sensitivity. The PhysiScore AUC for any IVH was 0.8092, whereas SNAP-II, SNAPPE-II, and CRIB had AUCs of 0.6761, 0.6924, and 0.7508, respectively. PhysiScore did not identify three infants who had severe IVH (grade 3 or 4) in the absence of any other HM complications. However, most infants who developed IVH can be found on the left side of the ROC, suggesting that PhysiScore offers high sensitivity without significant compromise in specificity (see Figure 5.4).

5.4.5 Importance of physiological features

Ablation analysis (comparison of model performance when different subsets of risk factors are included) was used to examine the contribution of score subcomponents in predicting HM versus LM. As expected, gestation and birth weight alone achieved reasonable predictive performance (AUC 0.8517). However, these two characteristics are not sufficient for individual risk prediction [Tyson *et al.*, 2008]. Notably, physiological parameters alone were more predictive than laboratory values alone (AUC, 0.8540 versus 0.7710, respectively). Adding

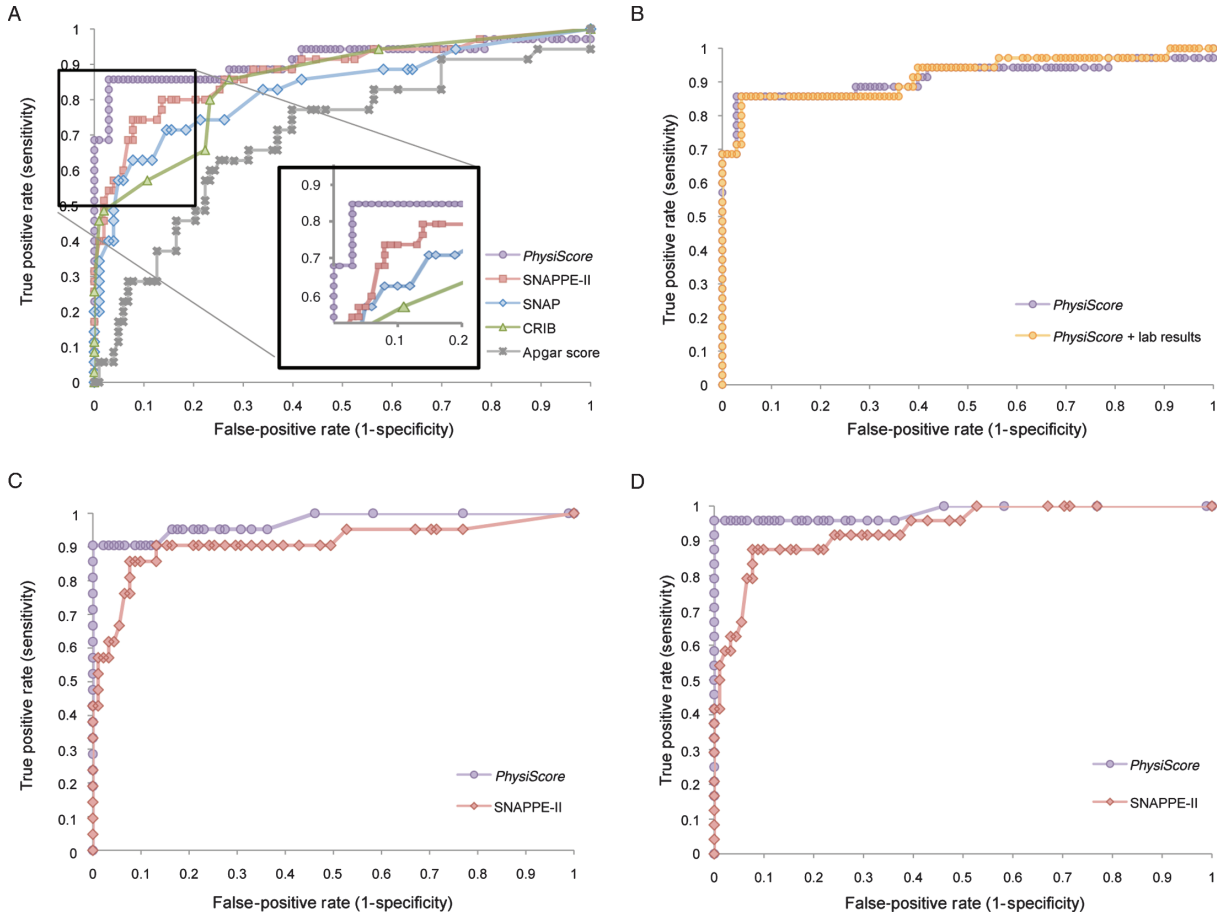


Figure 5.3: (A) ROC curves demonstrating PhysiScores performance in predicting high morbidity as it relates to conventional scoring systems. (B) PhysiScores performance with laboratory studies. (C) Predictions for infants with infection-related complications. (D) Predictions for infants with major cardiopulmonary complications.

physiological parameters to gestation and birth weight (that is, PhysiScore) increased the AUC to 0.9129, a significantly ($P < 0.01$) [DeLong *et al.*, 1988] better prediction than gestation and birth weight alone. Addition of laboratory values and physiologic characteristics did not significantly increase the AUC (0.9197), again suggesting that these parameters are redundant with the laboratory data in morbidity prediction.

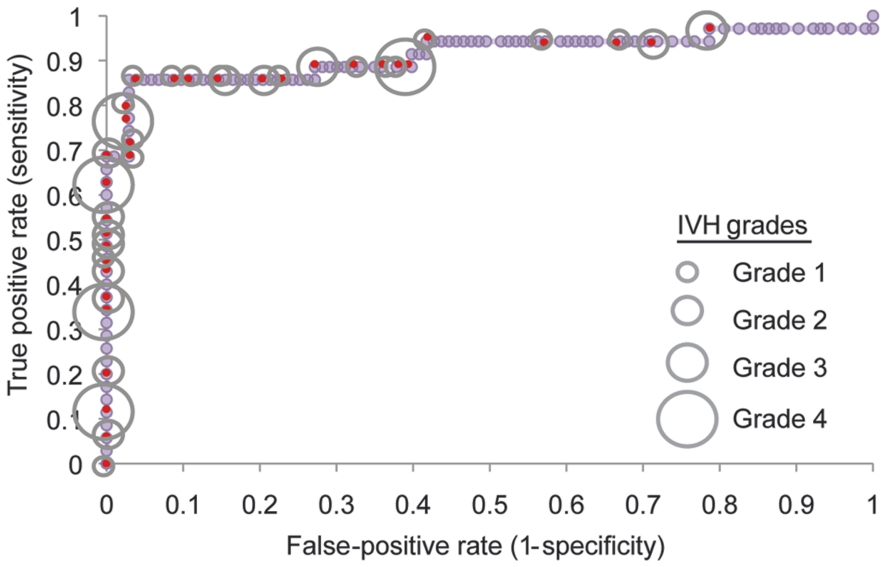


Figure 5.4: ROC curve demonstrating the limited sensitivity of PhysiScore in predicting morbidity for infants with IVH. Each circle represents the IVH grade of a preterm neonate overlaid on the respective score.

Examination of the learned weights (w_i in Eq. 5.1) of individual physiological parameters incorporated into PhysiScore (Figure 5.5A) demonstrated that short-term heart and respiratory rate variability made a significant contribution to the value of the PhysiScore, but long-term variability did not. Thus, short-term variability patterns (often difficult to see by eye, but easily calculated by PhysiScore) carried significant physiological information that long-term variability patterns did not.

Only three categories of commonly obtained physiological measurements were required for PhysiScore: heart rate, respiratory rate, and oxygen saturation. From these measures, using Bayesian modeling, we obtained individual curves that convey the probability of HM associated with individually calculated physiological parameters (Figure 5.5B).

As expected, a respiratory rate between 35 and 75 breaths per minute had a greater probability of being associated with health, whereas higher or lower rates carried a greater

probability of morbidity. A decreased short-term heart rate variability also indicated increased risk, consistent with previous findings linking this parameter to sepsis [Williams and Galerneau, 2003a].

This visual analysis of the nonlinear relationships seen in Figure 5.5B also suggests unexpected associations. Short-term respiratory rate variability, not commonly used as a physiological marker, was associated with increased morbidity risk. Unlike residual heart rate variability, its effect was nonmonotonic. Risk curves describing oxygen saturation suggest, respectively, that risk increases significantly with mean saturations less than 92% and prolonged time spent ($> 5\%$ total time) at oxygen saturations below 85%. Oxygenation is routinely manipulated by physician intervention, suggesting that intervention failure (for example, the inability to keep saturations in a specific range) that allows desaturations lasting for $> 5\%$ of total time is associated with higher morbidity risk, a threshold that can now be prospectively assessed in clinical trials.

5.5 Discussion

We have developed a risk stratification method that predicts morbidity for individual preterm neonates by integrating multiple continuous physiological signals from the first 3 hours of life. This score is analogous to the Apgar score [Casey *et al.*, 2001], in that only physiological observations are used to derive morbidity and mortality predictions. However, the use of time series data combined with automated score calculation yields significantly more information about illness severity than is provided by the Apgar score.

5.5.1 Discriminative capacity

Past efforts have resulted in several illness severity scores that use laboratory studies and other perinatal data to achieve improved discriminative ability over the Apgar score alone. For all of the available neonatal illness scores, much of the discriminative ability comes from gestational age and birth weight. Nevertheless, it is well-recognized that age- and weight-matched neonates may have significantly different morbidity profiles [Tyson *et al.*, 2008]. The CRIB score uses logistic regression to define six factors and their relative weights in predicting mortality: birth weight, gestational age, congenital malformation, maximum base deficit in the first 12 hours, plus minimum and maximum FiO_2 (fraction of inspired oxygen) in the first 12 hours [Network, 1993]. SNAP-II and SNAPPE-II were both derived from

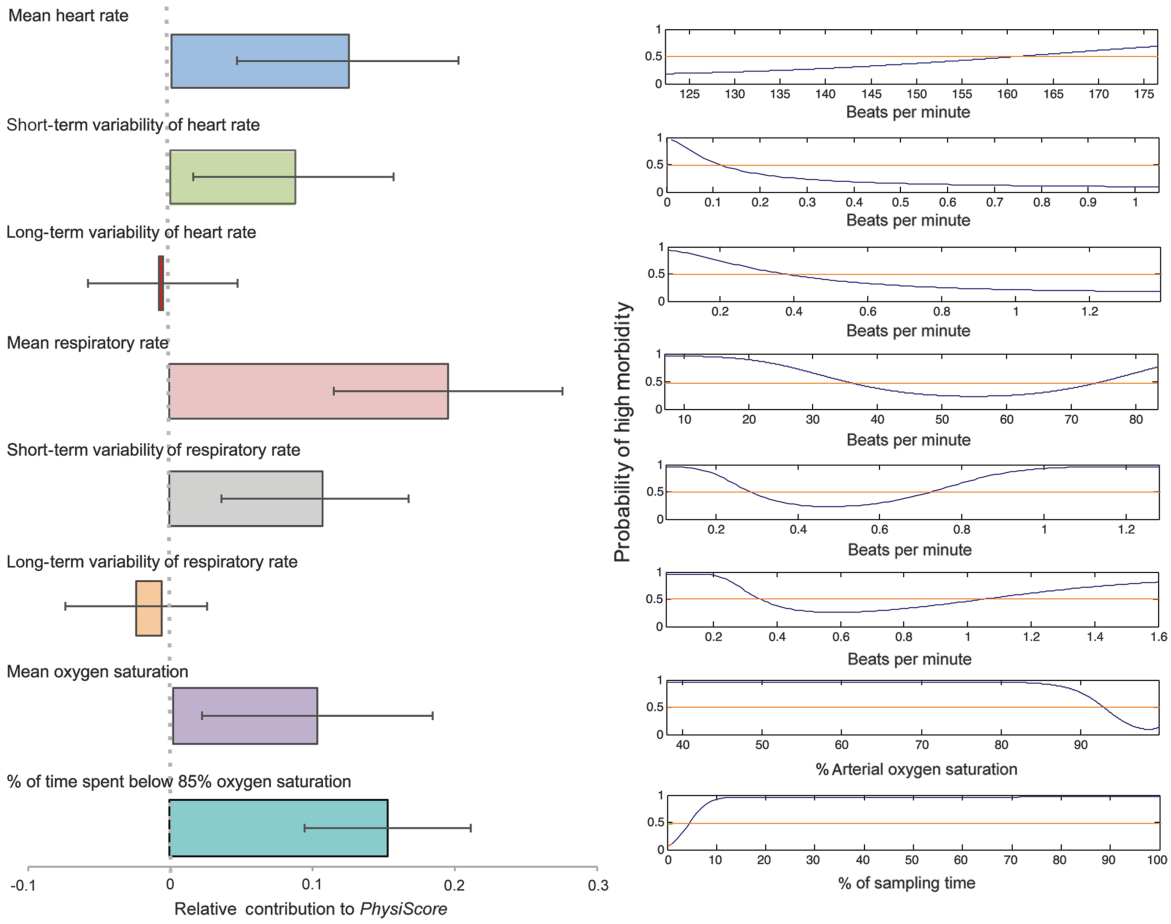


Figure 5.5: The significance of different physiological parameters in predicting high morbidity. (A) The learned weight (w_i in Eq. 5.1) for each physiological parameter incorporated in PhysiScore; error bars indicate variation in the weight over the different folds of the cross-validation. (B) The nonlinear function associating the parameter with the risk of high versus low morbidity.

SNAP. SNAP uses 34 factors identified by experts as important in the first 24 hours of life (specific laboratory data, minimum and maximum vital sign values, and other clinical signs). The resulting score correlated well with birth weight, mortality, length of stay, nursing acuity, and physician estimates of mortality, but was complex to calculate [Richardson *et al.*, 1993]. Logistic regression performed on the 34 factors in SNAP identified six variables most predictive of mortality that were recorded in the first 12 hours of life (lowest mean

blood pressure, lowest core body temperature, lowest serum pH, multiple seizures, urine output, and FiO2/PaO2 ratio); these were retained in SNAP-II. SNAPPE-II is calculated with the same data as SNAP-II, along with the 5-min Apgar score, small for gestational age status, and birth weight. The additional variables present in SNAPPE-II were found to be independent risk factors for mortality [Richardson *et al.*, 2001]. None of these scores, however, discriminate morbidity risk as well as PhysiScore, which integrates a small set of continuous physiological measures calculated directly from standard vital sign monitors.

An intriguing aspect of our findings is that PhysiScore provides high-accuracy predictions about morbidity risk from limited initial data (only 3 hours), even when such outcomes manifest days or weeks later (for example, BPD or NEC). Thus, it suggests that Physiscore may be reflecting an infant's predisposition to sickness. Furthermore, infants weaker at birth are most likely to develop major complications. PhysiScore gives positive weight to loss of short-term heart rate variability, much in the way that fetal heart rate monitoring uses loss of short-term heart rate variability to predict fetal distress and guide delivery management [Williams and Galerneau, 2003b]. PhysiScore additionally identifies short-term respiratory variability as having high predictive value, suggesting that further exploration of this factor in other settings might be warranted. Although the precise source of variability loss – either pre- or postnatally – is unknown, autonomic dysregulation likely plays a role.Whether short-term variability loss causes morbidity or is simply a marker of illness is not clear at this point.

Unlike fetal heart rate monitoring or heart rate spectral analysis [Tuzcu *et al.*, 2009], our approach uses multiple physiological parameters to improve accuracy and provide long-term predictions that extend beyond acute risk. Unlike biomarkers, such predictions are made with data that are already being collected in NICUs. Patient oxygenation, heart rate, and respiratory rate can be automatically processed to compute a score, and a predetermined sensitivity/specificity threshold can be used to make morbidity predictions to guide clinical actions, thereby removing the need for end-user expertise. When integrated into a bedside monitor, the algorithm would indicate the statistical likelihood that an individual patient is at high risk of major morbidities, allowing real-time use of the PhysiScore calculation. This method of deployment would effectively provide an automated electronic Apgar score, with significantly higher predictive accuracy regarding neonatal morbidity.

The PhysiScore's ability to assess physiologic disturbances before it can be confounded by medical intervention makes it highly descriptive of initial patient acuity; thus, it is

well suited as a tool for quality assessment between NICUs. Identification of a patient's future risk of developing HM complications may be useful for decision-making in primary nurseries to make more informed decisions regarding aggressive use of intensive care, need for transport to higher levels of care, and resource allocation. Such economic, social, and medical advantages should be evaluated in a large-scale clinical trial.

5.5.2 Technical considerations

Although we have a relatively small sample size, analysis methods appropriate to small sample sizes [Rangayyan, 2005] were used, and ROC curves were made only for morbidities seen in $> 10\%$ of our population. Our model, with its automatic factor modeling and selection, requires essentially no parameter tuning, which greatly helps to prevent overfitting in small samples.

In addition, our sample is from a single tertiary care center and was limited to patients born in our institution to ensure that continuous physiological data were available for the first hours of life. Validation in other settings will be required.

Detection of IVH remains elusive in the field of neonatal medicine. Previous work reported that fractal analysis of the original newborn heartbeat may be an early indicator of IVH [Tuzcu *et al.*, 2009], but yielded no better sensitivity than PhysiScore. This study included 10 newborn very low birth weight infants with intraventricular hemorrhage (5 grade IV, 4 grade III, and 1 grade II) and 14 control infants without intraventricular hemorrhage. Performance of 70% sensitivity and 79% specificity was achieved in their study as compared to 80% sensitivity and 76% specificity for Physiscore. It is possible that the underlying pathophysiology of IVH is variable [McCrea and Ment, 2008], particularly in infants in whom severe IVH is the only morbidity. Although IVH is usually associated with cardiopulmonary instability, recent literature suggests that there may be genetic predisposition to isolated IVH, potentially limiting the role of antecedent physiological signals before large hemorrhages [Vannemreddy *et al.*, 2010]. Thus, it is possible that the small number of infants with isolated IVH that were not identified as high risk by PhysiScore represent a distinct subpopulation.

5.5.3 Advanced computational techniques in modern medical settings

The use of computer-based techniques to integrate and interpret patterns in patient data to automate morbidity prediction has the potential to improve medical care. The current

U.S. governmental mandate to improve electronic health record use and gain economic benefit from using digital data [111th United States Congress, 2009b] facilitates the use of computer based tools. Flexible Bayesian modeling with almost no tunable parameters allows our approach to be applicable to a range of different prediction tasks, allowing use of the highly informative but underused data obtained daily for thousands of acutely ill patients.

Subjects (N)	138
Birth weight (g)	1367 ± 440
Gestational age (weeks)	29.8 ± 3
Gender, female	68
Apgar score at 5 min	7 ± 3
SGA (≤5th percentile)	7
Multiple gestation	
Total	46
Twins	20
Triplets	6
Respiratory distress syndrome	112
Pneumothorax	10
Bronchopulmonary dysplasia	
Total	29
NOS*	2
Mild	12
Moderate	5
Severe	10
Pulmonary hemorrhage	2
Pulmonary hypertension	3
Acute hemodynamic instability	11
Retinopathy of prematurity (ROP)	
Total	25
Stage I	9
Stage II	12
Stage III	4
Intraventricular hemorrhage (IVH)	
Total	34
Grade 1	19
Grade 2	7
Grade 3	3
Grade 4	5
Posthemorrhagic hydrocephalus	6
Culture-positive sepsis	11
Necrotizing enterocolitis	
Total	8
Stage 1	2
Stage 2	4
Stage 3	2
Expired	4

Table 5.1: Baseline and disease characteristics of the study cohort. (SGA, small for gestational age; NOS, not otherwise specified.)

	Apgar	SNAP-II	SNAPPE-II	CRIB	PhysiScore
Predicting high morbidity	0.6978	0.8298	0.8795	0.8509	0.9151
Infection	0.7412	0.8428	0.9087	0.8956	0.9733
Cardiopulmonary	0.7198	0.8592	0.9336	0.9139	0.9828

Table 5.2: Performance summary with AUCs.

Chapter 6

Conclusion

Developing a deeper understanding of human health and disease is one of the most fundamental and fascinating problem we face. With widespread digitization of complete patient encounters and its availability through the electronic health record, we now have access to comprehensive and highly granular data regarding the movement of patients — their symptoms, interventions, outcomes — through the health care system. By modeling these data well, we can better infer health trajectories of both individuals and populations. This would enable us to stage early interventions or optimize over the many available treatment options. For example, subtle early disease signatures in the data such as the ones presented earlier in this thesis can indicate that a patient should be sent to a higher level of care; predicting severity and relapses can help avoid expensive re-hospitalizations; and discovering sub-populations that evolve differently can lead to effective treatment differentiation. In addition, models that highlight which missing measurements would be most informative in uncovering the progression of health status could provide active decision support to doctors. These are only a few examples of the myriad opportunities that computational tools built on EHR data can enable.

Contrasted with data collected from a randomized control trial where the data collection process is hypothesis driven i.e., only a limited set of indicators are collected on a predefined patient population required to validate a specific hypothesis, EHR data has the advantage of being more comprehensive. On the other hand, as discussed previously, this data is much more challenging. It is confounded by observed and unobserved interventions, is high-dimensional, highly unstructured, heterogeneous, noisy and is often missing systematically. Furthermore, high-quality ground truth data is scarce both because obtaining it requires

expensive annotator time and that the notion of ground truth is not always well understood. Thus, extracting the full value of the existing data requires novel approaches.

This thesis is a foray into how observational patient data from the EHR can be harnessed for making novel clinical discoveries. For this, one requires access to patient outcome data — which patient has which complications. We present a method for automated extraction of patient outcomes from EHR data; our method shows how natural languages cues from the physicians notes can be combined with clinical events that occur during a patient's length of stay in the hospital to extract high quality annotations. In addition, we develop novel methods for exploratory analysis and structure discovery in bedside monitor data. This data forms the bulk of the data collected on any patient yet, it is not utilized in any substantive way post collection. We present methods to extract clinically motivated *shape* and *dynamic* signatures. Our analysis led us to a novel use of this data for risk prediction in infants. Using features automatically extracted from physiologic signals collected in the first 3 hours of life, we developed Physiscore, a tool that predicts infants at risk for major complications downstream. Physiscore is both fully automated and significantly more accurate than the current standard of care. It can be used for resource optimization within a NICU, managing infant transport to a higher level of care and parental counseling. Overall, this thesis illustrates how the use of machine learning for analyzing these large scale digital patient data repositories can yield new clinical discoveries and potentially useful tools for improving patient care.

There are numerous ways in which one can continue to pursue this fruitful direction of research. Physiscore presented only a snapshot view of an infant's health status at 3 hours of life; continuously estimating an infant's health trajectory can yield real-time opportunities for customizing intervention. Furthermore, refining these predictions to make inferences at the level of individual diseases and a disease's progression can provide physicians with more easily actionable information. Extending these to other patient populations requiring long term support (e.g., adults with chronic complications) offers another tremendous opportunities for care optimization, both due to the gravity of these problems (e.g., almost 1 in 5 adults have a chronic condition) and the amount of data being collected on these patients due to their repeated exposure to the health system.

Comparative effectiveness is another growing area of national importance. A recent article [Sutherland *et al.*, 2009] in the New England Journal of Medicine showed that only about 30% of the regional variation in spending is attributable to variations in individual

health and socioeconomic status. The remaining expenditures result from inefficiencies in the health care system — discretionary decisions by physicians that are influenced by the local availability of hospital beds, imaging centers and other resources, as well as by a payment system that rewards growth and higher utilization. Analysis of health trajectories with respect to various interventions can be used to inform clinical practice guidelines. Methods that model variations in populations and resources can be used to tune guidelines to individual regions. Moreover, deploying these guidelines via clinical decision support tools can help reduce variability in care.

Modeling data to infer health trajectories from any of these populations will require coping with numerous issues simultaneously. Non-stationarity is commonplace: e.g., bilirubin levels and heart rates in infants change as they mature, and disease progression impacts many measurements. The timescales on which even a single signal evolves can differ greatly; from the slow, circadian rhythm, to the rapid effects of bradycardia visible in the heart rate. Discovering these varying paces requires learning multi-resolution models. The observed measurements are taken in various formscontinuous sensor outputs, lab results, and qualitative narratives from nurses and doctors. To integrate these diverse sources, we need hybrid probabilistic models that scale to large data. The high-dimensionality of the data requires intelligent methods of transfer learning (to leverage data across related tasks) and close attention to parsimony. And importantly, we need flexible ways of encoding bias from clinicians, who already understand the data better than anyone. We will increasingly have access to measurements at multiple levels of granularity, from the genetic level all the way up to physiologic signals and behavioral observations. While we cannot hope to model this deep stack at once, we can build rich models of subsystems, and incrementally grow and combine them to reach new insights about our physiology and better cues for the above-mentioned clinical tasks.

Large EMR repositories can also be used for validating our models retrospectively; this will speed up both the rate at which we are able to conduct new trials and their chance of success. For this, we must develop methods [Rosenbaum and Rubin, 1983] that can account for confounding factors ubiquitous in observation data. Conversely, by learning models of interventions and their effects on the physiologic subsystems, we may even be able to theoretically posit how clinical trials would progress with a varied combination of interventions, and prioritize trials based on outcome and cost.

We are at a very exciting time – the accelerating adoption of EHRs is creating vast

quantities of unexplored longitudinal health data. I believe the ability to effectively harness this data will revolutionize the quality of our healthcare system through early diagnoses and better-optimized care.

Bibliography

[111th United States Congress, 2009a] 111th United States Congress. *The American Recovery and Reinvestment Act of 2009*. Government Institutes/Bernan Press, 2009.

[111th United States Congress, 2009b] 111th United States Congress. *The American Recovery and Reinvestment Act of 2009 (Public Law 111-5) official text*. Government Institutes/Bernan Press, 2009.

[Aleks *et al.*, 2009] N. Aleks, S. Russell, M. Madden, K. Staudenmayer, M. Cohen, D. Morabito, and G. Manley. Probabilistic detection of short events with application to critical care monitoring. In *Neural Information Processing Systems (NIPS)*, 2009.

[Bar-Shalom and Fortmann, 1987] Y. Bar-Shalom and T.E. Fortmann. *Tracking and Data Association*. Academic Press Professional, Inc., 1987.

[Behrman and Butler, 2007] R. Behrman and A. Butler. *Preterm Birth: Causes, Consequences and Prevention*. National Academies Press, 2007.

[Blackwell and MacQueen, 1973] D. Blackwell and J.B. MacQueen. Ferguson distributions via polya urn schemes. In *Annals of Statistics, 1(2)*, 1973.

[Blei *et al.*, 2003] David Blei, Andrew Ng, and Michael Jordan. Latent Dirichlet allocation. In *Journal of Machine Learning Research*, 2003.

[Burton *et al.*, 2008] M.M. Burton, L. Simonaiis, and G. Schadow. Medication and indication linkage: A practical therapy for the problem list? In *American Medical Informatics Association (AMIA) annual symposium*, 2008.

[Campbell and Payne, 1994] J. Campbell and T. Payne. A comparison of four schemes for codifications of problem lists. In *Proceedings of the Annual Symposium in Computer Applications in Medical Care*, 1994.

[Casey *et al.*, 2001] B. M. Casey, D. D. McIntire, and K. J. Leveno. The continuing value of the apgar score for the assessment of newborn infants. New England Journal of Medicine, 2001.

[Chiu *et al.*, 2003] B. Chiu, E. Keogh, and S. Lonardi. Probabilistic discovery of time series motifs. In *Knowledge Discovery and Datamining (KDD)*, 2003.

[Crammer *et al.*, 2007] K. Crammer, M. Dredze, K. Ganchev, P.P. Talukdar, and S. Carroll. Automatic code assignment to medical text. In *Proceedings of the Workshop on BioNLP*, 2007.

[Crawshaw *et al.*, 2010] A.P. Crawshaw, C.J. Wotton, D.G.R. Yeates, M.J. Goldacre, and L.-P. Ho. Evidence for association between sarcoidosis and pulmonary embolism from 35-year record linkage study. In *Thorax*, 2010.

[D. Heckerman and Nathwani, 1992] E. Horvitz D. Heckerman and B. Nathwani. Toward normative expert systems: Part i. the pathfinder project. 1992.

[DeLong *et al.*, 1988] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. In *Biometrics*, 1988.

[Denton, 2005] A. Denton. Kernel-density-based clustering of time series subsequences using a continuous random-walk noise model. In *International Conference on Data Mining*, 2005.

[Diebel, 2008] J. Diebel. Bayesian Image Vectorization: the probabilistic inversion of vector image rasterization. Phd thesis, Computer Science Department, Stanford University, 2008.

[Ehrenkranz *et al.*, 2005] R. A. Ehrenkranz, M. C. Walsh, B. R. Vohr, A. H. Jobe, L. L. Wright, A. A. Fanaroff, L. A. Wrage, and K. Poole. National institutes of child health and human development neonatal research network, validation of the national institutes of health consensus definition of bronchopulmonary dysplasia. In *Pediatrics 116*, 2005.

[Ferguson, 1973] T.S. Ferguson. A bayesian analysis of some nonparametric problems. The Annals of Statistics, 1(2), 1973.

[Field and Grigsby, 2002] M. Field and J. Grigsby. Telemedicine and remote patient monitoring. In *Journal of American Medical Association*, 2002.

[Fine *et al.*, 1x998] S. Fine, Y. Singer, and N. Tishby. The Hierarchical Hidden Markov Model: Analysis and applications. In *Journal of Machine Learning (JMLR)*, 1x998.

[Fink and Gandhi, 2010] E. Fink and H. Gandhi. Compression of time series by extracting major extrema. In *Journal of Experimental and Theoretical AI*, 2010.

[for the Classification of Retinopathy of Prematurity, 2005] International Committee for the Classification of Retinopathy of Prematurity. The international classification of retinopathy of prematurity revisited. In *Arch. Ophthalmol.*, 2005.

[Fox *et al.*, 2007] Emily Fox, Erik Sudderth, Michael Jordan, and Alan Willsky. The sticky HDP-HMM: Bayesian nonparametric hidden Markov models with persistent states. Technical Report P-2777, MIT LIDS, 2007.

[Fox *et al.*, 2008] Emily Fox, Erik Sudderth, Michael Jordan, and Alan Willsky. Nonparametric Bayesian learning of switching linear dynamical systems. In *Neural Information Processing Systems (NIPS)*, 2008.

[Fox *et al.*, 2009] E. Fox, E. Sudderth, M. Jordan, and A. Willsky. Sharing features among dynamical systems with Beta Processes. In *Neural Information Processing Systems (NIPS)*, 2009.

[Friedman *et al.*, 1998] J.H. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. In *Technical Report, Dept. of Statistics, Stanford University*, 1998.

[Friedman *et al.*, 2004] C. Friedman, L. Shagina, Y. Lussier, and G. Hripcsak. Automated encoding of clinical documents based on natural language processing. In *Journal of the American Medical Informatics Association*, 2004.

[Gallier, 1999] J. Gallier. *Curves and Surfaces in Geometric Modeling: Theory and Algorithms*. Morgan Kaufmann, 1999.

[Gandhi *et al.*, 2011] T.K. Gandhi, G. Zuccotti, and T.H. Lee. Incomplete care – on the trail of flaws in the system. In *New England Journal of Medicine*, 2011.

[Gelman *et al.*, 1995] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian data analysis*. Chapman and Hall, 1995.

[Goldacre *et al.*, 2009] M.J. Goldacre, C.J. Wotton, and D.G.R. Yeates. Cancer and immune-mediated disease in people who have had meningococcal disease: record-linkage studies. In *Epidemiol. Infect. 137(5)*, 2009.

[Goldberger *et al.*, 2000] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. In *Circulation 101(23)*, 2000.

[Goldstein *et al.*, 2007] I. Goldstein, A. Arzumtsyan, and O. Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. In *American Medical Informatics Association (AMIA) annual symposium*, 2007.

[Griffin *et al.*, 2005a] M. Griffin, Douglas Lake, Eric Bissonette, Frank Harrell, T. O'Shea, and J. Moorman. Heart rate characteristics: novel physiomarkers to predict neonatal infection and death. In *Pediatrics*, 2005.

[Griffin *et al.*, 2005b] M. P. Griffin, D. E. Lake, and J. R. Moorman. Heart rate characteristics and laboratory tests in neonatal sepsis. In *Pediatrics*, 2005.

[Hastie *et al.*, 2001] T. Hastie, R. Tibshirani, and J. H. Friedman. The elements of statistical learning: Data mining, inference, and prediction. Springer, 2001.

[Hausdorff *et al.*, 2000] J.M. Hausdorff, A. Lertratanakul, M.E. Cudkowicz, A.L. Peterson, D. Kaliton, and A.L. Goldberger. Dynamic markers of altered gait rhythm in amyotrophic lateral sclerosis. In *Journal of Applied Physiology*, 2000.

[Himes *et al.*, 2009] B.E. Himes, Y. Dai, I.S. Kohane, S.T. Weiss, and M.F. Ramoni. Prediction of chronic obstructive pulmonary disease (copd) in asthma patients using electronic medical records. In *Journal of American Medical Informatics Association*, 2009.

[Höppner, 2002] F. Höppner. Knowledge discovery from sequential data. 2002.

[Ishwaran and Zarepour, 2000] H. Ishwaran and M. Zarepour. Markov chain monte carlo in approximate dirichlet and beta twoparameter process hierarchical models. In *Biometrika, 87*, 2000.

[Ishwaran and Zarepour, 2002a] H. Ishwaran and M. Zarepour. Dirichlet prior sieves in nite normal mixtures. In *Statistica Sinica 12*, 2002.

[Ishwaran and Zarepour, 2002b] H. Ishwaran and M. Zarepour. Exact and approximate sum-representation for the Dirichlet process. In *Canadian Journal of Statistics*, 2002.

[Jelinek, 1997] F. Jelinek. *Statistical Method for Speech Recognition*. MIT Press, 1997.

[Joachims, 2006] T. Joachims. Training linear SVMs in linear time. In *Knowledge Discovery and Datamining (KDD)*, 2006.

[Kass and Steffey, 1989] R. Kass and D. Steffey. Approximate bayesian inference in conditionally independent hierarchical models (parametric empirical bayes models). In *Journal of American Statistical Association*, 1989.

[Keogh and Folias, 2002] E. Keogh and T. Folias. UCR time series data mining archive. 2002.

[Keogh *et al.*, 2000] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra. Dimensionality reduction for fast similarity search in large time series databases. In *Journal of Knowledge and Information Systems*, 2000.

[Kim *et al.*, 2006] S. Kim, P. Smyth, and S. Luther. Modeling waveform shapes with random effects segmental hidden Markov models. In *Journsl of Machine Learning Research*, 2006.

[Kliegman and Walsh, 1987] R. M. Kliegman and M. C. Walsh. Neonatal necrotizing enterocolitis: Pathogenesis, classification, and spectrum of illness. In *Curr. Probl. Pediatr.*, 1987.

[Knaus *et al.*, 1985] W. A. Knaus, E. A. Draper, D. P. Wagner, and J. E. Zimmerman. Apache ii: a severity of disease classification system. In *Critical care medicine*, 1985.

[Koller and Friedman, 2009] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[Kurihara *et al.*, 2007] K. Kurihara, M. Welling, and Y.W. Teh. Collapsed variational Dirichlet process mixture models. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.

[Lafferty *et al.*, 2001] L. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference in Machine Learning (ICML)*, 2001.

[Lee *et al.*, 2009] Honglak Lee, Yan Largman, Peter Pham, and Andrew Ng. Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Neural Information Processing Systems (NIPS)*, 2009.

[Li *et al.*, 2008] L. Li, H.S. Chase, and C. et al. Patel. Comparing icd9-encoded diagnoses and nlp-processed discharge summaries for clinical trials pre-screening: A case study. In *American Medical Informatics Association (AMIA) annual symposium*, 2008.

[Liao *et al.*, 2007] L. Liao, D.J. Paterson, D. Fox, and H.A. Kautz. Learning and inferring transportation routines. International Joint Conference on Artificial Intelligence (IJCAI), 2007.

[Liao *et al.*, 2010] K.P. Liao, T. Cai, V. Gainer, S. Goryachev, Q. Zeng-Treitler, S. Raychaudhuri, P. Szolovits, S. Churchill, S. Murphy, and I. et. al. Kohane. Electronic medical records for discovery research in rheumatoid arthritis. In *Arthritis Care Research*, volume 62, 2010.

[Listgarten *et al.*, 2005] J. Listgarten, R. Neal, S. Roweis, and A. Emili. Multiple alignment of continuous time series. In *Neural Information Processing Systems (NIPS)*, 2005.

[Liu *et al.*, 2001] H. Liu, Y. A. Lussier, and C. Friedman. Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method. In *Journal of Biomedical Informatics*, 2001.

[McCrea and Ment, 2008] H. J. McCrea and L. R. Ment. The diagnosis, management, and postnatal prevention of intraventricular hemorrhage in the preterm neonate. In *Clinical Perinatology*, 2008.

[Melton and Hripcsak, 2005] G.B. Melton and G. Hripcsak. Automated detection of adverse events using natural language processing of discharge summaries. In *American Medical Informatics Association (AMIA) annual symposium*, 2005.

[Meystre *et al.*, 2008] S.M. Meystre, G.K. Savova, K.C. Kipper-Schuler, and J.F. Hurdle. Extracting information from textual documents in the ehr: A review of recent research. In *IMIA Yearbook of Medical Informatics*, 2008.

[Mietus *et al.*, 2000] J.E. Mietus, C.-K. Peng, P.Ch. Ivanov, and A.L. Goldberger. Detection of obstructive sleep apnea from cardiac interbeat interval time series. In *Comput Cardiol*, 2000.

[Minnen *et al.*, 2007] D. Minnen, C. L. Isbell, I. Essa, and T. Starner. Discovering multivariate motifs using subsequence density estimation and greedy mixture learning. In *American Association of Artificial Intelligence*, 2007.

[Moody and Mark, 2001] G.B. Moody and R.G. Mark. The impact of the MIT-BIH Arrhythmia Database. 2001.

[Mueen *et al.*, 2009] A. Mueen, E. Keogh, Q. Zhu, S. Cash, and B. Westover. Exact discovery of time series motifs. In *Siam Conference on Data Mining*, 2009.

[Network, 1993] The International Neonatal Network. The crib (clinical risk index for babies) score: A tool for assessing initial risk and comparing performance of neonatal intensive care units. In *Lancet*, 1993.

[Nyquist, 1928] H. Nyquist. Certain topics in telegraph transmission theory. In *AIEE*, 1928.

[Oates, 2002] T. Oates. PERUSE:an unsupervised algorithm for finding recurring patterns in time series. In *International conference on Data Mining*, 2002.

[Pakhomov *et al.*, 2005] S. Pakhomov, J. Buntrock, and P.H. Duffy. High throughput modularized nlp system for clinical text. In *Association for Computational Linguistics*, 2005.

[Papile *et al.*, 1978] L. A. Papile, J. Burstein, R. Burstein, and H. Koffler. Incidence and evolution of subependymal and intraventricular hemorrhage: A study of infants with birth weights less than 1,500 gm. In *Journal of Pediatrics*, 1978.

[Peng *et al.*, 1995] C.-K. Peng, S. Havlin, H.E. Stanley, and A.L. Goldberger. Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. In *Chaos 5*, 1995.

[Petri *et al.*, 2010] H. Petri, D. Maldonato, and N.J. Robinson. Data-driven identification of co-morbidities associated with rheumatoid arthritis in a large us health plan claims database. In *BMC Musculoskeletal Disorders 11(1)*, 2010.

[Poon and Merrill, 1997] C.-S. Poon and C.K. Merrill. Decrease of cardiac chaos in congestive heart failure. In *Nature*, 1997.

[Poritz, 2009] A.B. Poritz. Linear predictive hidden markov models and the speech signal. In *Symposium on Applications of Hidden Markov Models to Test and Speech*, 2009.

[Quinn *et al.*, 2009] J. Quinn, C. Williams, and N. McIntosh. Factorial switching linear dynamical systems applied to physiological condition monitoring. In *IEEE Trans. Pattern Analysis Machine Intelligence*, 2009.

[Ramage *et al.*, 2009] D. Ramage, D. Hall, Nallapati R, and C. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2009.

[Rangayyan, 2005] R. M. Rangayyan. *Biomedical Image Analysis*. Biomedical Engineering Series (CRC Press), 2005.

[Reis *et al.*, 2009] B.Y. Reis, I.S. Kohane, and K.D. Mandl. Longitudinal histories as predictors of future diagnoses of domestic abuse: modelling study. In *BMJ*, 2009.

[Richardson and Green, 1997] S. Richardson and P. J. Green. On bayesian analysis of mixtures with an unknown number of components. In *Journal of the Royal Statistical Society. Series B*, 1997.

[Richardson *et al.*, 1993] D. Richardson, J. Gray, M. McCormick, K. Workman, and D. Goldmann. Score for neonatal acute physiology: a physiologic severity index for neonatal intensive care. In *Pediatrics*, 1993.

[Richardson *et al.*, 2001] D. K. Richardson, J. D. Corcoran, G. J. Escobar, and S. K. Lee. Snap-ii and snappe-ii: Simplified newborn illness severity and mortality risk scores. In *Journal of Pediatrics*, 2001.

[Robertson *et al.*, 1992] P. A. Robertson, S. H. Sniderman, R. K. Laros Jr., R. Cowan, D. Heilbron, R. L. Goldenberg, J. D. Iams, and R. K. Creasy. Neonatal morbidity

according to gestational age and birth weight from five tertiary care centers in the united states, 1983 through 1986. In *American Journal of Obstetric Gynecology*, 1992.

[Rosenbaum and Rubin, 1983] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. In *Biometrika 70*, 1983.

[Ross, 2004] S. M. Ross. Introduction to probability and statistics for engineers and scientists. Elsevier Academic Press, 2004.

[Saria *et al.*, 2010] S. Saria, A. Rajani, J. Gould, D. Koller, and A. Penn. Integration of early physiological responses predicts later illness severity in preterm infants. In *Science Trans. Med.*, 2010.

[Schnabel *et al.*, 2009] R. B. Schnabel, L. M. Sullivan, D. Levy, M. J. Pencina, J. M. Massaro, R. B. DAgostino Sr., C. Newton-Cheh, J. F. Yamamoto, J. W. Magnani, T. M. Tadros, W. B. Kannel, T. J. Wang, P. T. Ellinor, P. A. Wolf, R. S. Vasan, and E. J. Benjamin. Development of a risk score for atrial fibrillation (framingham heart study): A community-based cohort study. In *Lancet*, 2009.

[Schulte-Frohlinde *et al.*, 2002] V. Schulte-Frohlinde, Y. Ashkenazy, A. L. Goldberger, P. Ivanov, M. Costa, A. Morley-Davies, H. E. Stanley, and L. Glass. Complex patterns of abnormal heartbeats. Phys. Rev. E Stat. Nonlin. Soft Matter Phys., 2002.

[Schwarz, 1978] G. Schwarz. Estimating the dimension of a model. In *Annals of Statistics*, 1978.

[Sethuraman, 1994] J. Sethuraman. A constructive definition of Dirichlet priors. In *Statistics Sinica*, 1994.

[Shumway, 1988] R. Shumway. *Applied statistical time series analysis.* Prentice Hall, 1988.

[Solt *et al.*, 2009] I. Solt, D. Tikk, V. Gl, and Z.T. Kardkovcs. Semantic classification of diseases in discharge summaries using a contextaware rule-based classifier. In *Journal of the American Medical Informatics Association*, 2009.

[Solti *et al.*, 2008] I. Solti, B. Aaronson, and B. Fletcher et. al. Building an automated problem list based on natural language processing: Lessons learned in the early phase of development. In *AMIA Annual Symposium Proceedings*, 2008.

[Stacey and McGregor, 2007] M. Stacey and C. McGregor. Temporal abstraction in intelligent clinical data analysis: A survey. In *AI in Medicine*, 2007.

[Sutherland *et al.*, 2009] J.M. Sutherland, E.S. Fisher, and J.S. Skinner. Getting past denial — the high cost of health care in the united states. In *New England Journal of Medicine*, 2009.

[Syed *et al.*, 2009a] Z. Syed, P. Indyk, and J. Guttag. Learning approximate sequential patterns for classification. In *Journal of Machine Learning Research (JMLR)*, 2009.

[Syed *et al.*, 2009b] Z. Syed, B. Scirica, S. Mohanavelu, P. Sung, C. Cannon, Peter Stone, Collin Stultz, and John Guttag. Relation of death within 90 days of non-st-elevation acute coronary syndromes to variability in electrocardiographic morphology. In *American Journal of Cardiology*, 2009.

[Teh *et al.*, 2006] Y.W. Teh, M.I. Jordan, M. Beal, and D. Blei. Hierarchical Dirichlet processes. In *Journal of American Statistical Association*, 2006.

[Tsuji *et al.*, 1994] H. Tsuji, F. J. Venditti Jr., E. S. Manders, J. C. Evans, M. G. Larson, C. L. Feldman, and D. Levy. Reduced heart rate variability and mortality risk in an elderly cohort. the framingham heart study. In *Circulation*, 1994.

[Tuzcu *et al.*, 2009] V. Tuzcu, S. Nas, U. Ulusar, A. Ugur, and J. R. Kaiser. Altered heart rhythm dynamics in very low birth weight infants with impending intraventricular hemorrhage. In *Pediatrics*, 2009.

[Tyson *et al.*, 2008] J. E. Tyson, N. A. Parikh, J. Langer, C. Green, and R. D. Higgins. National institute of child health and human development neonatal research network, intensive care for extreme prematurity moving beyond gestational age. In *New England Journal of Medicine*, 2008.

[Ueda *et al.*, 1998] N. Ueda, R. Nakano, Z. Ghahramani, and G. Hinton. Split and merge EM algorithm for improving Gaussian mixture density estimates. In *Neural Information Processing Systems (NIPS)*, 1998.

[Uzuner, 2009] O. Uzuner. Recognizing obesity and co-morbidities in sparse data. In *Journal of the American Medical Informatics Association*, 2009.

[Vannemreddy *et al.*, 2010] P. Vannemreddy, C. Notarianni, K. Yanamandra, D. Napper, and J. Bocchini. Is an endothelial nitric oxide synthase gene mutation a risk factor in the origin of intraventricular hemorrhage? In *Neurosurg. Focus*, 2010.

[Viterbi, 1967] A.J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. In *IEEE Transactions on Information Theory*, 1967.

[Wang and McCallum, 2006] X. Wang and A. McCallum. Topics over Time: A non-Markov continuous time model of topical trends. In *Knowledge Discovery and Datamining (KDD)*, 2006.

[Wang *et al.*, 2008] C. Wang, D. Blei, and D. Heckerman. Continuous time dynamic topic models. In *Uncertainty in Artificial Intelligence (UAI)*. 2008.

[Whitlock *et al.*, 2009] G. Whitlock, S. Lewington, P. Sherliker, R. Clarke, J. Emberson, J. Halsey, N. Qizilbash, R. Collins, and R. Peto. Body-mass index and cause-specific mortality in 900 000 adults: Collaborative analyses of 57 prospective studies. In *Lancet*, 2009.

[Williams and Galerneau, 2003a] K. P. Williams and F. Galerneau. Intrapartum fetal heart rate patterns in the prediction of neonatal acidemia. In *American Journal of Obstetric Gynecology*, 2003.

[Williams and Galerneau, 2003b] K. P. Williams and F. Galerneau. Intrapartum influences on cesarean delivery in multiple gestation. In *Acta Obstet. Gynecol. Scand.*, 2003.

[Williams *et al.*, 2005a] C. Williams, J. Quinn, and N. McIntosh. Factorial switching Kalman filters for condition monitoring in neonatal intensive care. In *Neural Information Processing Systems (NIPS)*, 2005.

[Williams *et al.*, 2005b] C. Williams, J. Quinn, and N. McIntosh. Factorial switching Kalman filters for condition monitoring in neonatal intensive care. In *Neural Information Processing Systems (NIPS)*, 2005.

[Zhu and Hastie, 2004] J. Zhu and T. Hastie. Classification of gene microarrays by penalized logistic regression. In *Biostatistics*, 2004.