Representation Learning in Multi-dimensional Clinical Timeseries for Risk and Event Prediction

by

Marzyeh Ghassemi

Submitted to the Department of Electrical Engineering and Computer Science in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2017

Author . . . .
Department of Electrical Engineering and Computer Science
May 19, 2017

Certified by . . . .
Peter Szolovits
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by . . . .
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Representation Learning in Multi-dimensional Clinical Timeseries for Risk and Event Prediction

by

Marzyeh Ghassemi

Submitted to the Department of Electrical Engineering and Computer Science
on May 19, 2017, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computer Science and Engineering

Abstract

There are major practical and technical barriers to understanding human health, and therefore a need for methods that thrive on large, complex, noisy data. In this work, we present machine learning methods that distill large amounts of heterogeneous health data into latent state representations. These representations are then used to estimate risks of poor outcomes, and response to intervention in multivariate physiological signals. We evaluate the reduced latent representations by 1) establishing their predictive value in important clinical tasks and 2) showing that the latent space representations themselves provide useful insight into underlying systems. In particular, we focus on case studies that can provide evidence-based risk assessment and forecasting in settings with guidelines that have not traditionally been data-driven.

In this thesis we evaluate several methods to create patient representations, and use these features to predict important outcomes. Representation learning can be thought of as a form of phenotype discovery, where we attempt to discover spaces in the new representation that are markers of important events. We argue that these latent representations are useful markers when they 1) create better prediction results on outcomes of interest, and 2) do not duplicate features that are currently known bio-markers.

We present four case studies of learning representations, and evaluate the representations on real predictive tasks. First, we create forward-facing prediction models using baseline clinical features, and those from a Latent Dirichlet Allocation (LDA) model trained with clinical progress notes. We then evaluate the per-patient latent state membership to predict mortality in an intensive care setting as time moves forward. Second, we use non-parametric Multi-task Gaussian Process (MTGP) hyper-parameters as latent features to estimate correlations within and between signals in sparse, heterogeneous time series data. We evaluate the hyper-parameters for forecasting missing signals in traumatic brain injury patients, and predicting mortality in intensive care unit patients. Third, we train switching-state autoregressive models (SSAMs) to model the underlying states that emit patient vital signs over time. We evaluate the time-specific latent state distributions as features to predict vasopressor onset and weaning in intensive care unit patients. Finally, we use statistical and symbolic features extracted from wearable ambulatory accelerometers (ACC) mounted to the neck to classify patient pathology, and stratify patients' risk of voice misuse. We evaluate the utility of both statistically generated features and symbolic representations of glottal pulses towards patient classification.

Thesis Supervisor: Peter Szolovits
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

I would like to acknowledge several people who were impacted in the making of this thesis.

I would like to thank my parents for their love and support. My father, Abbas Ghassemi, has always been convinced that I could easily do anything I wanted to. However, I would like to point out that I have my PhD now, and the joke still is not funny. My mother was the best teacher, mentor and friend I could have asked for, and is still the only person who can remind me that things will be fine.

My husband and children have been incredible during my PhD. Eric Benjamin Munson might not have thought I was serious when I told him over a decade ago that living with me was going to involve a lot of traveling, but I think we've established I was. He has always been kind, funny, and utterly without roadrage — which makes him the perfect spouse for the greater Boston area. Raziyeh Elise Munson-Ghassemi has lived on the MIT campus since she was 8 months old and it shows. I'll be asking her future partner if they've read her papers. Abbas Benjamin Munson-Ghassemi has found a way to jump off of every structure on campus, and many off campus. I can only hope any new place we live will offer fewer jumping points. Somayeh Marit Munson-Ghassemi has kindly waited until after I defended before joining us.

My friends have made living in this cold, snowy, windy city much more fun than I thought possible. Many thanks go to Jen "Barefoot Lake Hiking" Gong, Tristan "Heirloom Tomato" Josef Naumann, Maggie "I Can See Your Hair Through The Glass" Makar, Layla "I Love You Enough To Wear A Pink Dress" Shaikhley, Rhonda "Serial Killers Have Traffic Violations" Shafaei, Zainab "Nanvaei in Gringolandia" Hosseini, Anna "The Nature" Haggman, and Michelle "The Nature Eats Your Flesh" Padua.

Finally, I would like to thank my advisor Peter Szolovits and my thesis committee members John Guttag and Leo Anthony Celi. Without their support, guidance, and humor, this experience would have been substantially poorer.

This doctoral thesis has been examined by a Committee of the Department of Electrical Engineering and Computer Science as follows:

Doctor Leo Anthony Celi ....

Member, Thesis Committee
Attending Physician at Beth Israel Deaconess Medical Center
Principal Research Scientist, Massachusetts Institute of Technology
Assistant Clinical Professor, Harvard University School of Public Health

Professor John Guttag ....

Member, Thesis Committee
Professor of Electrical Engineering and Computer Science

Professor Peter Szolovits ....

Thesis Supervisor
Professor of Electrical Engineering and Computer Science

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Patient health is managed by the flow of information, but the volume of such records can make it difficult for care-staff to identify the information relevant to patient care. In a fragmented healthcare system, accurate knowledge of a patient's disease state is critical. Modern electronic healthcare records contain an increasingly large amount of data including high-frequency signals from biomedical instrumentation, intermittent results from lab tests, and text from notes.

Good physicians are able to sift through large amounts of often-irrelevant data in order to discover information relevant to a patient's current underlying state. In the same vein, a goal of clinical inference and prediction is often to stratify patients who are similar in some underlying, hidden, characteristics. In this thesis, our goal is to provide clinical staff with support tools that will help them make better decisions. We focus on evidence generation in three specific subtasks: early prediction of actionable in-patient interventions in the intensive care unit (ICU), identifying good representations for post-discharge outcome prediction, and creating evidence-based diagnosis of voice disorders in an out-patient setting.

1.1 Need for Evidence Generation with Electronic Health Records

Evidence generation in clinical settings is important because clinical decision-making is often made in settings of limited knowledge and high uncertainty. For example, only 10 of 72 common ICU interventions have been associated with improved long-term outcomes [73]. Further, randomized controlled trials (RCTs) are commonly used to generate evidence about medical practices, but do not cover a majority of treatments that are commonly used: only 10–20% of treatments are based on evidence from an existing RCT [64, 67].

Critical care in the United States costs more than $55 billion annually [38], and is rapidly growing as a specialty worldwide. Internationally, the mortality rate in the ICU is approximately 15–16% [104]. However, despite the economic and human impact of critical illness, practice in this field is informed by a relative paucity of high-quality trials [49].

Outside of critical care, most people spend a majority of their time outside of a clinical environment. Ambulatory monitoring provides an exciting avenue for detecting and managing illness in non-clinical settings — especially for chronic conditions such as cardiovascular disease [97], diabetes [76], or voice disorder [65]. As chronic conditions become more prevalent [72], it is increasingly important to understand how ambulatory monitoring can be used to affect structural behavior changes that can improve patients' lives.

In order to provide clinical staff with actionable recommendations for patient care, we plan to gain insight from healthcare data, focusing on ambulatory bio-monitors and the Electronic Health Record (EHR). Ambulatory monitoring is particularly compelling for conditions that are behaviorally induced, like some voice disorders [42]. As in other chronic conditions, patients with voice disorders are known to be poor judges of their behaviors [9,71,82]. Ambulatory monitoring allows for a more accurate understanding of such chronic conditions, but comes with the computational challenge of finding behaviors that may be harmful to an individual long-term from large amounts of high-frequency data. EHR data are also becoming more prevalent — EHR systems that meet federal requirements are present in most acute care hospitals (97% in 2014 [13]) and office-based physicians' practices (78% in 2015 [70]). This availability allows new investigations into evidence-based decision support for critical care, where we can learn when patients are at high risk for mortality or need a given intervention. However, working with such "secondary" data is difficult.

1.2 Challenges of Electronic Health Record Data

Unlike many other types of data used in clinical trials, secondary data gathered from EHRs and out-patient monitoring solutions are not gathered specifically to answer a hypothesis. Instead, their primary use case is to monitor a patient or make decisions about patient care. Thus, there is innate confounding by indication, because the data are gathered in response to the needs of the use case, which creates obvious interactions between the patient's condition and the data collection process. These conditions can lead to several issues that make learning difficult, primarily that the data are 1) *heterogeneous*, 2) *sparse*, and 3) full of *uncertainty.*

First, EHRs contain heterogeneous data types ranging from notes typed at different times during carestaff shift hours, to labs that are recorded when the clinicians order them, to vitals that are noted hourly (or more often), to more static demographic data. These large differences in data type, time scale, and sampling rates make modeling underlying physiology challenging. Second, each data type has a different type of sparsity; a vital may be *unmeasured* because a sensor fell off a patient, a lab value may be measured but the value *unreported* in the EHR, or there may be *no follow-up* when a patient is prescribed a medication to know if the medication was filled or (if filled) taken to completion. In the clinical domain, labels themselves are often weak proxies for underlying truth — a diagnostic clinical code for diabetes may indicate that a person has been diabetic for years, or that the clinician suspects there may be diabetes and cannot bill patient insurance for the test until there is an appropriate code in their EHR. There is also uncertainty in the bias of the presented data; clinical data is often recorded only when a patient is sick, leading to a biased sample of physiological state and history. Uncertainty also stems from the relative nature of clinical data, where the content of a statement can vary wildly depending on the larger context, e.g., “the patient improved” for a comatose patient versus a patient with a flu.

Given these issues, applying state of the art machine learning methods out-of-the-box without a deep understanding of the data and methodology is dangerous. This was illustrated recently by work that considered applying a neural network to predict the risk of dying of pneumonia in a hospital population [11]. The model learned that those patients with asthma had a lower risk of dying, which is clinically (and intuitively) incorrect. As noted by the authors, the “aggressive care received by asthmatic pneumonia patients (in the training set) was so effective that it lowered their risk of dying from pneumonia compared to the general population.”

1.3 Framework

Our main contributions are in two areas: 1) forming machine learning tasks that are clinically meaningful, and 2) emphasizing methods that create meaningful representations of clinical data. We present several examples of creating actionable clinical insights from machine learning, with a focus on representations that work for the nuances of clinical datasets.

First, when forming a clinically meaningful task, it is critical to understand the desired learning outcome. Forming machine learning tasks in a meaningful way with clinical data is more difficult than it may initially seem. For example, in the hospital setting, it may initially seem natural to predict mortality using all available data from a patient up until their time of death. However, such a task would lead to learning that patients die when their support machines are turned off in the preceding hour. Similarly, we may want to predict what type of procedures patients with given conditions are most likely to need during their stay. However, if we use a patient’s ICD9 codes as a proxy for their condition (e.g., as input) for predicting procedures (e.g., as output), we have inherently cheated in the prediction task because many “diagnostic” codes are created post-care (after a patient’s death or discharge) using the record of hospital care they experienced.

Second, when learning representations, we balance the need for predictive power with the desire to find representations appropriate for each presented problem. Representation learning can be characterized as finding representations of data that contain useful information towards some goal, and has been successfully used for tasks in speech recognition, object recognition, and natural language processing [4]. Dealing with representations explicitly may be advantageous because they can conveniently express general priors that are not specific to a single predictive task. With any representation method, the reduced state space can be viewed as a new set of biomarkers. In this work, we think of the latent "biomarkers" as attributes or features that can be learned over data from many patients, and that separate patients out into meaningful groups. Our goal is to create representations that are useful for important prediction tasks, and independent of the existing biomarkers known for a particular outcome.

In each piece of presented work, we were guided by three general representation goals. First, we want representations to be useful abstractions of data that disentangle underlying factors. E.g., we can learn a latent state representation for hourly vitals that maximizes the likelihood of observing the physiological data. Second, representations should enable semi-supervised learning of an outcome Y. For example, the temporal trends pulled out of a convolutional neural network should not only represent meaningless variation of a signal, but rather be useful for recognition of physiological decline. Third, use of the representation should be "shared" across many learning tasks (many Y's). For example, features that are able to predict the future need for several different types of interventions.

To achieve these goals, we focus on semi-supervised frameworks, where we learn a model of data in an unsupervised setting, but predict values for new cases in supervised settings. In general, discriminative learning is based on the observation of data drawn from some distribution. In unsupervised learning, this may correspond to observing points $x_1, \ldots, x_N$, and modeling the distribution the data comes from. In supervised learning, this may correspond to observing pairs of points $(x_1, y_1), \ldots, (x_N, y_N)$ and predicting a new y_{N+1} given an x_{N+1}.

1.3.1 Organization

The rest of this thesis is organized as follows:

1. **Early Prediction of In-Patient ICU Interventions**: We target representations that are predictive of early need for in-patient ICU interventions. The representations are learned first in an unsupervised way, and then applied to targeted supervised prediction tasks [35, 36].

2. **Representations for Post-Discharge Outcome Prediction**: We learn representations of clinical notes that are predictive of in-hospital mortality, post-discharge mortality, and psychiatric readmission [30, 32, 86].

3. **Voice Disorder Detection in Wearable Out-patient Devices**: We create features of out-patient ambulatory accelerometer signals that are able to distinguish between patients with two different types of voice disorder [33, 34].

1.3.2 Assumptions

There are many assumptions made in the work presented here. First, in Chapter 2 we use the actual behavior of clinicians as "correct" labels, even though they may not be. This "wisdom of the crowd" approach to treatment has been harmful in clinical situations. For example, clinicians initially thought that estrogen was cardioprotective because menopausal women had a higher incidence of coronary heart disease [28]. Based on this, hormone therapy was routinely prescribed as a preventative measure. However, subsequent large trials reported either no benefit, or an increase in adverse cardiac events like coronary heart disease [27], stroke and venous thromboembolism [78]. We also emphasize that our work presents learned *associations* rather than learned *mechanisms*. We learn attributes of a patient that are predictive of a targeted outcome, and attempt to relate our findings to known (or possible) mechanisms, However, we would ideally we would like to understand the mechanism (in this case, the pathophysiology) that has led to a patient's outcome. To address this issue, future work should target treatment comparisons using reinforcement learning or casual inference frameworks.

In Chapter 3, we use mortality at various intervals as a proxy for patient acuity. The underlying issue behind this assumption is the general lack of a quantitative measure of patient health. We use likelihood of mortality as a convenient proxy, but the correlation is not perfect. For example, patients who are not acutely ill may die of causes unrelated to their current or past severity of illness. Further, in a modern ICU patients may be kept "alive" for an extended amount of time regardless of their severity of illness.

In Chapter 4 we have targeted prediction of a diagnosed voice disorder that occurs without any recorded anatomical changes. In the absence of physical pathology, we rely on a high-level "disordered" label where subjects' instantaneous behaviors are producing voices that are difficult for the subject to maintain, or for others to understand. This focuses our work on identifying the changes between the accelerometer patterns recorded in the available populations: patients pre-treatment, patients post-treatment, and subjects without an established voice disorder. It is possible that our "control" population could eventually develop a voice disorder, or that the patterns we associate with disordered voices in our population would not be consistent in a larger population. To address these questions, larger datasets of individuals with a voice disorder should be studied, and there is value to identifying potential mid-level classification targets, e.g., classes or levels of disorder during specific segments of voicing.

Chapter 2

Early Prediction of In-Patient ICU Interventions

2.1 Background

Decision-making in the intensive care unit (ICU) requires responding quickly to rapidly changing situations, but the efficacy of many interventions remains unquantified [103, 105], while other interventions have been shown to be ineffective or harmful to patients [73]. The vast amounts of data that are collected in ICUs—vital signs, clinical notes, fluids, medications—suggest an opportunity for more data-driven decision-making. Many works have used these ICU measurements to predict in-hospital or 30-day mortality of patients in particular disease subgroups [10, 14, 30, 75]. However, these risk scores are of limited value to clinicians, who must make decisions of how and when to treat patients regardless of their underlying acuity.

This work takes an important step toward the *actionable* use of ICU data by modeling interventions in the ICU. We focus on vasopressors, a class of drug used to elevate mean arterial pressure. While vasopressors are commonly used in the ICU, few controlled clinical trials have documented improved outcomes from their use [68], and they may even be harmful in some populations [23]. We consider two important questions relating to vasopressor administration. First, we ask when a patient will require a vasopressor. Knowing who will need a vasopressor administration even a few hours in advance can help the clinical staff plan and execute interventions in a safer, more efficient manner. Second, we ask whether a patient currently on a vasopressor is ready to be weaned from it. In addition to being conservative about when the patient is ready for weaning, anecdotally clinicians report that patients are often left on interventions longer than necessary because the staff are attending to other patients. However, extended interventions are both costly and detrimental to patient health [23].

Unfortunately, making decisions from data generated in the ICU is challenging: clinical signals are often irregularly sampled and contaminated by interference and human error. Strong modeling assumptions are typically used to clean and impute the signals [50, 61]. However, these imputation techniques can introduce noise and bias into models [60]. They also generally do not account for the highly dependent temporal nature of the data [44, 53, 62]. Dynamical system models, which impute data by building a model of how the data evolve, provide an alternative to interpolation-based imputation techniques. In particular, switching-state autoregressive systems (SSAMs) have been used to impute signals, identify artifacts, and discover physiological states in a variety of critical settings [56, 79]. SSAMs are attractive methods for modeling physiologic signals because they express the notion that the dynamics of the physiologic signal will change depending on some internal patient health state; given a patient's health state, the set of physiological signals at the next time depends only on the current signals. This assumption considerably simplifies the training of the model, resulting in a more robust predictor. Interpretation of a SSAM is also relatively simple, because at each time, a patient is assigned to exactly one discrete hidden state, rather than some more complex embedding. Finally, we note that [46] demonstrated how an appropriate discretization of the physiological signal can improve performance on downstream tasks; as such our transition models are distinct from the SSAM in [56, 79].

2.2 Overview

Our work was done in close collaboration with Mike Wu and Finale Doshi-Velez of Harvard. We use switching state models to model the physiological state of the patient. Unlike prior work, we focus on actionable predictions regarding interventions, rather than mortality. We also consider a much higher-dimensional space of 83 different physiological signals. Unlike prior work, we focus on actionable predictions regarding interventions, rather than mortality. We also consider a high-dimensional space of physiological signals and make use of signal discretization to improve performance on downstream tasks. Specifically, we

- define three clinically actionable prediction tasks: immediate need for an intervention, need for an intervention in the near future, and when a patient is ready to be weaned from an intervention,
- achieve state-of-the-art predictions for both intervention-onset tasks using only physiological signals in a large, public ICU dataset,
- quantify unnecessary extra intervention time. To our knowledge, ours is the first study to use predictive models to address this question.

2.3 Data

The MIMIC II 2.6 database includes retrospective electronic medical records (EMRs) for 26,870 adult hospital admissions recorded between 2001 and 2008. [89] The creation and use of the MIMIC database was approved by the Institutional Review Boards of both BIDMC and MIT (IRB Protocol 2001-P-001699/3). Many ICU patients have a limited chance of survival, regardless of clinical intervention. Therefore, our cohort contains only adult patients on their first ICU stay without orders for reduced care (e.g., “comfort measures only,” “do not resuscitate,” “do not intubate,” or “CPR not indicated”). Following prior work [43], we also excluded patients with less than 12 hours of data or more than 96 hours of data to avoid a group of fundamentally sicker patients. These criteria allowed us to focus on situations in which clinical decisions might have a positive effect, rather than penalizing a classifier for situations where a patient is taken off life support. Applying these filters resulted in an initial cohort of 15,695 patients: 4,331 who were administered vasopressors (positive class) and 11,364 without vasopressors (control class).

2.4 Predictive Tasks

We consider three tasks: predicting (1) Imminent Vasopressor Need, (2) Short-term Vasopressor Need, and (3) Wean Readiness (see Figure 2-1 for an illustration).

1. Task 1: Imminent Vasopressor Need. We define imminent vasopressor need as requiring a vasopressor within the next 2 hours. For each patient, we make predictions every hour until the first vasopressor administration or the end of stay. We only predict the first vasopressor administration because patients with multiple vasopressors are likely to be in fundamentally different physiological situations.

2. Task 2: Short-term Vasopressor Need. We define short-term need if the patient is stable enough not to require vasopressor administration for the next 4 hours but will require vasopressor administration in the following 2 hours. Predicting who will require vasopressors in the near future—but not now—can help manage ICU logistics and ensure that the patient is ready for the intervention. We make hourly predictions until the first vasopressor administration or the end of stay.

3. Task 3: Wean Readiness. Vasopressors are administered via IV, and patients are weaned by gradually reducing the dose. We define Wean Readiness as being able to stop administration completely within 2 hours, and a successful wean as not requiring vasopressors again within 4 hours.

2.5 Feature Construction

Numeric trends are generally produced by the bedside monitors once per second, but often stored only once every 5 to 60 minutes. In this work we use the nurse-validated vital sign

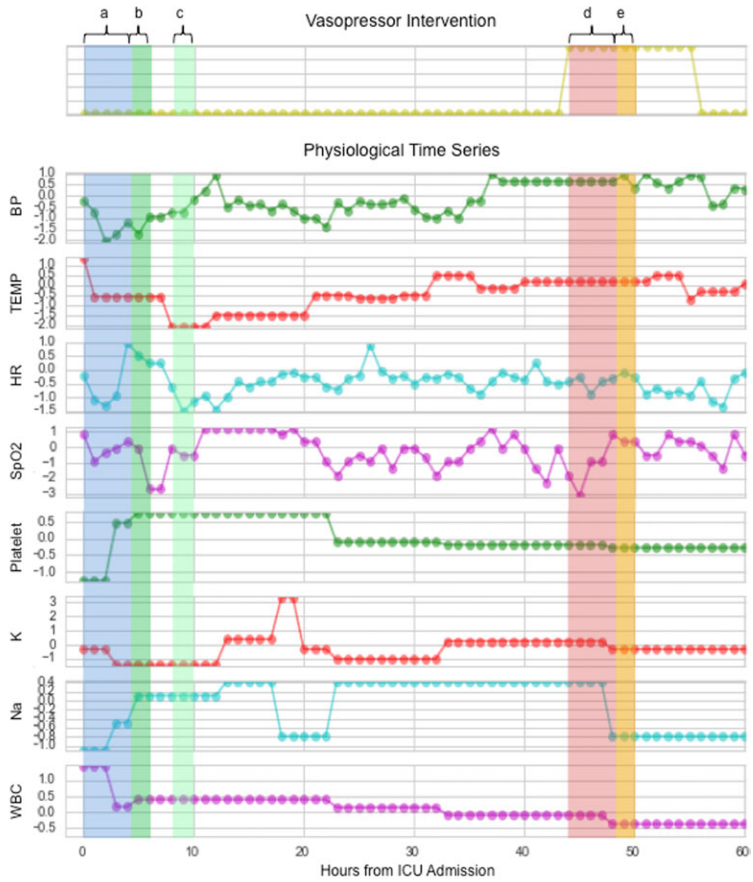


Figure 2-1: A subset of physiological timeseries with prediction windows highlighted. Predicting Imminent Vasopressor Need (Task 1) evaluates features from window “”**a**” on vasopressor need in window “**b**”. Predicting Short-term Vasopressor Need (Task 2) evaluates features from window **a** on vasopressor need in window “**c**”. Predicting Wean Readiness (Task 3) evaluates features from window “**d**” on the successful weaning of vasopressors in window “**e**”.

trends from the clinical information system, which are most-often sampled on an hourly basis. Variables were discretized using the mean and standard deviation from the training set. In other work, variables were discretized along “normal” value ranges using clinical knowledge [46]. In this work, we discretized values by rounding per-variable z-scores to integer values in -4:4; we added an extra value for missing values so each new physiological variable took on 10 discrete values. The additional value used to indicate *missing* was specifically not interpreted as an ordinal, as each of the discretized values was viewed as a possible emitted “character”. This discretization procedure helps the model to avoid fitting to small variations in the physiological signal and to identify global structure in the data while respecting the missing data rather than imputing it. Vasopressor administration variables were post-processed to recover continuous segments of administration.

2.5.1 Extracting and Processing Signals

Data were extracted from the MIMIC II 2.6 database. Data was gathered from four ICUs at the Beth Israel Deaconess Medical Center (BIDMC): medical (MICU), surgical (SICU), coronary care unit (CCU), and cardiac surgery recovery unit (CSRU).

We considered a total of 18 physiological variables for our model. These were the time series of 7 nurse-verified vital signs: heart-rate (HR), mean arterial blood pressure (MEAN BP), blood oxygenation level (SPO2), temperature (TEMP), spontaneous respiration rate (RESP), first inspired oxygen (FIO2), and urine output (URINE); the time series of 11 laboratory measurements: blood urea nitrogen (BUN), hematocrit (HCT), creatinine (CREAT), bicarbonate (BICAR), lactate (LACT), magnesium (Mg), potassium (K), sodium (Na), glucose (GLU), platelet count (PC), and white blood cell count (WBC); and 9 static variables: admitting age, gender, first SAPS I score, first SOFA acuity score, first weight, first ICU service type, body mass index (BMI), use of pacemaker, and whether the patient was noted as "at risk" for falls.

We first binned into hours from when the patient was admitted. If there were multiple values indicated for time series variables, the value for that hour was the mean of the values noted. To handle missing data, we only incorporated the 10 total features with greater than 10% non-missing entries (MEAN BP, TEMP, HR, SPO2, FIO2, RR, GLU, BICAR, HCT, K), and smoothed the data through sample-and-hold. All 9 static variables were included, yielding a total set of 19 physiological variables.

Extracting and Processing Outcomes We extracted vasopressor administration as any medication event with a generic or brand-name vasopressor label, including dopamine, epinephrine, isuprel, levophed, vasopressin, and neosynephrine. We considered any modification of vasopressor settings to be a binary indicator of vasopressor administration in the hour it occurred in. Because continuing vasopressor administration is not always noted in the electronic health record, we interpolated any vasopressor gaps less than 4 hours as being continuously on the medication, unless there was an explicit stoppage of the medication noted. From this smoothed timeseries, we computed the start time of the i^{th} vasopressor administration $t_n^{v_i}$ and its corresponding wean $t_n^{w_i}$ for each patient n.

Basic Statistics Table 2.1 shows the mean for features of patients before and during vasopressor administration, as well as patients who never had vasopressor administered during their stay. The means of the intervention and control groups are largely similar, except for the variables corresponding to which ICU the patient is in (MICU, SICU, CCU, CSRU, and FICU) and the risk of falls (a proxy for frailty).

2.5.2 Feature Overview

For each task being evaluated at hour t of patient n, we considered three types of input features: (1) Raw, (2) SSAM and (3) Combined. The raw features are the previous 4 hours of multidimensional z-scored physiological data at hour t of patient n, appended with the seven static admissions features. We learn the SSAM (switching state autoregressive model) features in an unsupervised fashion using the Raw Features. The Combined Features were obtained by concatenating the raw and SSAM feature vectors. (Figure 2-2)

2.5.3 SSAM

The physiological signals (Raw Features) $\boldsymbol{x}_t^n$ of a patient n at time t form a vector in $\mathbb{R}^D$ of D measurements, some of which may be missing. For each patient n, we observe a sequence of $\{\boldsymbol{x}_1^n, \boldsymbol{x}_2^n, \boldsymbol{x}_3^n, \ldots \boldsymbol{x}_{T_n}^n\}$ of length T_n. We train a switching state autogressive model to learn a hidden sequence of discrete, scalar variables $\{y_1^n, y_2^n, y_3^n, \ldots y_{T_n}^n\}$, that determine the transition dynamics of the observed variables $\boldsymbol{x}_t^n$. These variables y_t^n can be interpreted as the *physiological state* of the patient.

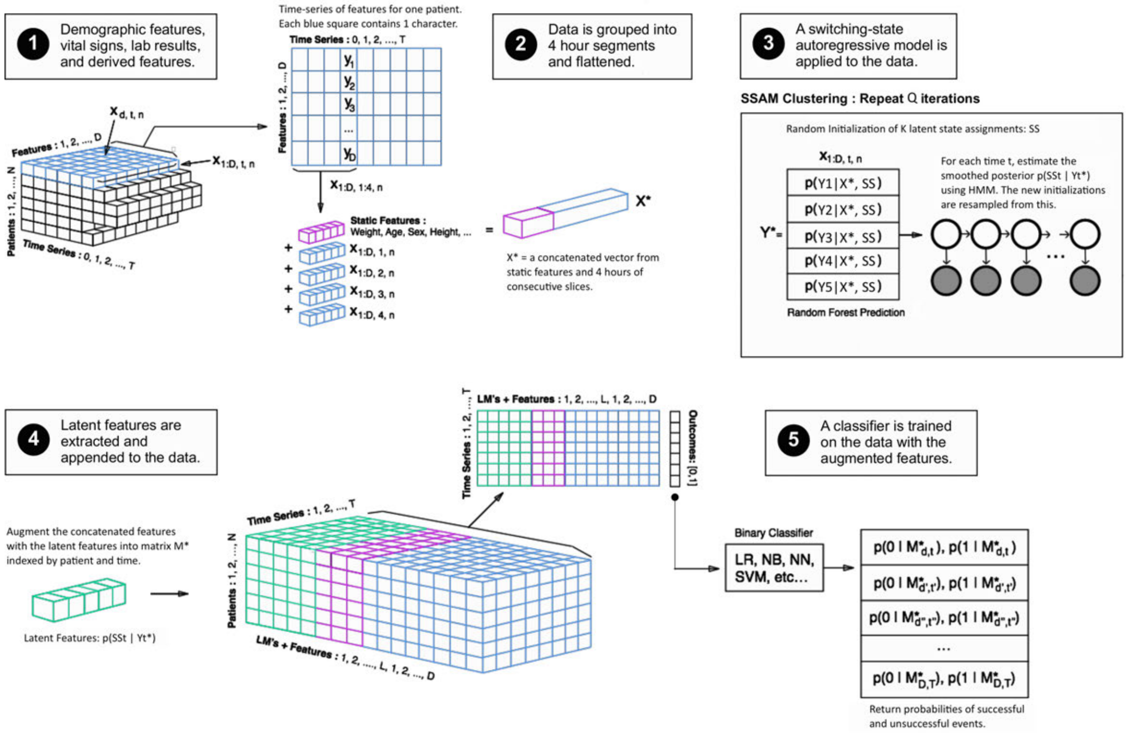

1
Demographic features, vital signs, lab results, and derived features.
Time-series of features for one patient. Each blue square contains 1 character.
Time Series : 0, 1, 2, ..., T
Features : 1, 2, ..., D
y1
y2
y3
...
yD
2
Data is grouped into 4 hour segments and flattened.
3
A switching-state autoregressive model is applied to the data.
x d, t, n
Features : 1, 2, ..., D
x 1:D, t, n
Patients : 1, 2, ..., N
Time Series : 0, 1, 2, ..., T
x 1:D, 1:4, n
Static Features : Weight, Age, Sex, Height, ...
+ x 1:D, 1, n
+ x 1:D, 2, n
+ x 1:D, 3, n
+ x 1:D, 4, n
=
X*
X* = a concatenated vector from static features and 4 hours of consecutive slices.
SSAM Clustering : Repeat Q iterations
Random Initialization of K latent state assignments: SS
x 1:D, t, n
Y* =
p(Y1|X*, SS)
p(Y2|X*, SS)
p(Y3|X*, SS)
p(Y4|X*, SS)
p(Y5|X*, SS)
Random Forest Prediction
For each time t, estimate the smoothed posterior p(SSt | Yt*) using HMM. The new initializations are resampled from this.
...
4
Latent features are extracted and appended to the data.
Augment the concatenated features with the latent features into matrix M* indexed by patient and time.
Latent Features: p(SSt | Yt*)
Time Series : 1, 2, ..., T
Patients : 1, 2, ..., N
LM's + Features : 1, 2, ..., L, 1, 2, ..., D
LM's + Features : 1, 2, ..., L, 1, 2, ..., D
Time Series : 1, 2, ..., T
Outcomes: [0,1]
5
A classifier is trained on the data with the augmented features.
Binary Classifier
LR, NB, NN, SVM, etc...
p(0 | M*d,t), p(1 | M*d,t)
p(0 | M*d',t'), p(1 | M*d',t')
p(0 | M*d'',t''), p(1 | M*d'',t'')
...
p(0 | M*D,T), p(1 | M*D,T)
Return probabilities of successful and unsuccessful events.

Figure 2-2: Overall graphical flow of experiment. (1) Baseline demographic features (e.g. age, sex, etc.), vital signs (heart rate, temperature, blood pressure, etc.), lab results (glucose levels, bicarbonate levels, etc.), and derived features (BMI) are extracted from the database for a filtered selection of patients. (2) For each patient, corresponding vital signs and lab results data are grouped into 4 hour blocks, flattened, and appended to the demographic features. (3) A switching-state autoregressive model is used the model the time series. (4) Latent features are then defined as the probability of each state at each time; these are appended to the features from step (2). (5) Given these features, a classifier is trained to predict the outcome of interest (e.g. vasopressor administration).

We assume that the transition dynamics of the hidden sequence $\{y_t^n\}$ follow a Markov model. Specifically, let there be K possible discrete hidden states. Conditioned on the current physiological state y_t^n, the distribution over the next state is given by

$$y_t^n \sim f_y(\cdot|y_{t-1}^n)$$

where the elements of the transition function f_y can be compactly represented in a $K \times K$ matrix. We place a non-uniform prior over the transition matrix to reflect a bias toward a patient staying in the same physiological state for extended periods of time, and we place a prior π_y over the initial state y_0^n.

Given the physiological state sequence, the observations are generated by an autoregressive model indexed by the hidden state y_t^n for each dimension d:

$$\boldsymbol{x}_t^n(d) \sim f_x(\cdot|\boldsymbol{x}_{t-1}^n \ldots \boldsymbol{x}_{t-M}^n, \theta_{d,y_{t-1}^n}) \tag{2.1}$$

where θ refers to the parameters of the transition model and M is the number of previous states that we consider when making predictions for the next state. Importantly, we assume that each dimension d has its own, independent transition function. This modeling choice is reasonable because the measurements come from many different types of variables which can be expected to behave with different dynamics. We considered several options for the autoregressive transition model f_x, but settled on using random forest classifiers. As with the hidden state sequence, we assume a prior distribution $\pi_{\boldsymbol{x}}$ for the initial measurement $\boldsymbol{x}_0^n$ (note that the hidden state y_0^n only governs the choice of transition dynamics, not the output itself).

The likelihood of the model is given by

$$L(\{y\}, \{\theta\}|\{\boldsymbol{x}\}) \;\; = \;\; \prod_{n}^{N} \pi_y(y_0^n)\pi_{\boldsymbol{x}}(\boldsymbol{x}_0^n) \prod_{t=1}^{T_n} f_y(y_t^n|y_{t-1}^n) \prod_{d=1}^{D} f_x(\boldsymbol{x}_t^n(d)|\boldsymbol{x}_{t-1}^n, \theta_{d,y_{t-1}^n})$$

where T_n is the number of observations for patient n.

Our model contains two sets of latent variables: the hidden physiological state sequences for each patient $\{y_1^n, y_2^n, y_3^n, \ldots y_{T_n}^n\}$ and the transition parameters $\theta_{d,k}$ for each measurement dimension d and physiological state k. Our inference alternates between updating each of these sets of variables. Inference was run for 45 iterations, starting with a random assignment of states to y_t^n. Tempering was used to used to avoid local optima [29].

2.5.4 Updating Autoregressive Function Parameters

Given the hidden state sequences for the patients $\{y_1^n, y_2^n, y_3^n, \ldots\}$, we can split the patient data into K sets of tuples $\{(\boldsymbol{x}_{t-M}^n, \ldots, \boldsymbol{x}_{t-1}^n, \boldsymbol{x}_t^n)\}$ for which $y_{t-1}^n = k$. In other words, we split the cases into subsets that share the same previous state. The transition dynamics of each of these K sets are distinct; for each of these sets, we train D classifiers, one for each dimension. The input to the classifier is the previous sequence $(\boldsymbol{x}_{t-M}^n, \ldots, \boldsymbol{x}_{t-1}^n)$ and the output is the $\boldsymbol{x}_t^n(d)$.

In equation 2.1, we used $\theta_{k,d}$ to denote the parameters of this classifier. Any classifier that can output the probability of a measurement $p(\boldsymbol{x}_t^n(d)|\boldsymbol{x}_{t-1}^n, \ldots, \boldsymbol{x}_{t-M}^n)$ can be used, and the training process will depend on the particular choice of classifier, where any given classifier will often have standard implementations in many machine learning libraries. Importantly, once the subsets have been formed, the training of the classifiers can be parallelized across each physiological state k and each output dimension d. For the autoregressive models, we considered random forests (with 10 trees).

2.5.5 Updating Physiological State and Transition Parameters

Given the autogressive transition parameters $\{\theta_{p,k}\}$ and the transition function f_y, we can update the state sequences using the standard forward-backward algorithm for HMMs [81]. We use a variant called forward-filtering backward sampling in which we first recursively compute the probabilities of each state y_t^n given the data $\{\boldsymbol{x}_1^n, \boldsymbol{x}_2^n, \ldots, \boldsymbol{x}_t^n\}$ up to time t:

$$\begin{aligned} Pr(y_t^n|\{\boldsymbol{x}_1^n, \boldsymbol{x}_2^n, \ldots, \boldsymbol{x}_t^n\}, \{y_1^n, y_2^n, \ldots, y_{t-1}^n\}, \{\theta_{p,k}\}) & \\ \propto f_y(y_t^n|y_{t-1}^n) \prod_{d=1}^{D} T_x(\boldsymbol{x}_t^n(d)|\boldsymbol{x}_{t-1}^n, \theta_{p,y_{t-1}^n}) & \\ \cdot\, Pr(y_{t-1}^n|\{\boldsymbol{x}_1^n, \boldsymbol{x}_2^n, \ldots, \boldsymbol{x}_{t-1}^n\}, \{y_1^n, y_2^n, \ldots, y_{t-2}^n\}, \{\theta_{p,k}\}) & \quad (2.2) \end{aligned}$$

and then sampling each state y_t^n in a backwards pass:

$$y_t^n \sim f_y(y_t^n|y_{t+1}^n)\; Pr(y_t^n|\{\boldsymbol{x}_1^n, \boldsymbol{x}_2^n, \ldots, \boldsymbol{x}_t^n\}, \{y_1^n, y_2^n, \ldots, y_{t-1}^n\}, \{\theta_{p,k}\})$$

where the final state $y_{T_n}^n$ is simply sampled from equation 2.2.

Given the hidden state sequence $\{y_1^n, y_2^n, y_3^n, \ldots\}$, we learn the transition function f_y by sampling from a Dirichlet distribution with parameters set by posterior transition counts. For the SSAM, we used 5 hidden states and a diagonal transition matrix $T(y, y')$ with $T(y, y) = 0.8$ and the remaining entries $T(y, y') = 0.05$ for $y \neq y'$.

2.6 Evaluation Procedure and Model Settings

The window size for autoregressive model (M) was empirically determined. We found that having $M > 1$ decreased performance because then the training data was too sparse. During training, we also explored a range of gap sizes for prediction (up to 12 hours), but found that it was significantly harder to predict outcomes further away. When selecting the window size for features, we found that features collected over 8 hours did not perform significantly differently from those over 4 hours. We believe this may be due to the 8 previous hours not providing more relevant temporal information given the previous 4 hours. We experimented with using up to 10 states in our model, but found that there was no significant difference in predictive performance; this may be due to the specific predictive tasks we have chosen in this work.

For each task (administration and weaning) we trained the SSAM on the patients from the positive class only. For vasopressor administration, we used all time points up to the administration of the vasopressor. For weaning, we only considered data immediately after the start of administration (control class) and immediately before the wean (positive class). At time t we computed the probabilities of being in each SSAM state over the last 4 hours for all patients and all times in our cohort and used those as input features. Because there are k states at every hour, 4 hours of previous data creates $4k$ SSAM Features.

Models were built from each of these features using three different classifiers: a linear-kernel support vector machine (SVM), naive Bayes (NB), and L2-regularized logistic regression (LR). Standard packages and settings were used for the SVMs, NB, and LR classifiers. All analysis was performed in Python 2.7.

2.7 Results

2.7.1 Predicting Vasopressor Administration Improved by SSAM Features

Table 2.2 compares the performance of all feature sets on Tasks 1 and 2 (imminent and short-term administration prediction) using L2-regularized logistic regression averaged over five repetitions. The LR classifiers tended to have the best prediction performance across feature sets; the results with all classification methods can be found in the Appendix.

Simply using the global SSAM features gives an area under the receiver operating curve (AUC) of 0.87 (± 0.009) for imminent need prediction, and 0.83 (± 0.008) for short-term need prediction. The combined features achieve the best results, and consistently improve AUCs over only using the raw features - AUCs of 0.92 (± 0.002) and 0.88 (± 0.006) for imminent need and short-term need prediction respectively.

2.8 Predicting Vasopressor Weaning Improved by SSAM Features

Following the best results from administration prediction, we trained a classifier for Task 3 to predict successful weaning on those patients who were alive 30 days post-discharge. We focus on this longer-term survival group in order to distinguish between physiological patterns that lead to successful weans in patients who were able to survive all aspects of their hospital treatment. The raw features obtained an AUC of 0.67 (± 0.008), SSAM(NB) features were AUC 0.63 (± 0.021), and Raw+SSAM(NB) features were AUC 0.71 (± 0.005).

2.8.1 Quantifying Unnecessary Intervention Time Prior to a Wean

Our quantitative results above discriminate situations in which the clinician may have attempted to wean too early, causing the wean to be unsuccessful. However, clinicians report that patients are often left on interventions for much longer than necessary. We focus on the first time that our classifier predicted a successful wean for each patient in Task 3, and examine the difference in time between these predicted weaning times and actual weaning times. As shown in Figure 2-3, a significant portion of patients were successfully weaned at the right time, but the heavy tail depicted suggests that many patients suffered from extended interventions.

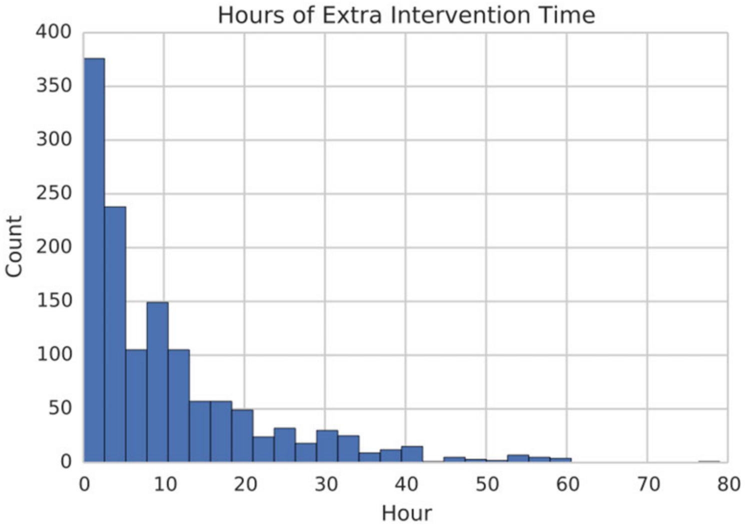


Figure 2-3: Histogram of excess time for which patients could have been successfully weaned according to the classifier.

We choose two patients from different points in the histogram in figure 2-3 and examine their medical notes.

Case 1: Figure 2-4 shows our probability of a successful wean from the time on vasopressor onset for a 72 year old man with coronary artery disease who was put on mechanical ventilation and vasopressors while in the ICU. The probability of a successful wean is low while the patient fails mechanical ventilation weaning early on in his stay, and immediately post-extubation. It is explicitly noted in his record at the lowest probability of wean that the patient is dependent on the vasopressors he is receiving. The patient stabilizes as the probability of wean success climbs, and the clinical staff actually begin to wean the patient near the highest predicted success in our estimates.

Case 2: Figure 2-5 shows a similar plot for a 62 year old male patient with a cardiac catheterization. The probability of successful wean remains low while the patient is given a course of treatment and fluids, and he struggles with a low central venous pressure (CVP) and increasing hematocrit (HCT). When the nursing staff notes an increasing need for vasopressors, the corresponding probability of a wean dips further. During recovery, our modelâĂŹs improved wean success matches the nurseâĂŹs note that the patient should be weaned in the following day. In this case, the wean happens almost 10 hours after our model predicts that it could successfully have been done. However, this is likely due to clinical staff schedules, which vary widely in the ICU. For legal and ethical reasons, there is also a bias to maintain interventions in ICU patients rather than withdraw too early, even if a patient seems to be stable.

2.8.2 Clinical Relevance of Discovered States

The previous sections show that our SSAM features improve our ability to predict vasopressor administration and weaning. We theorize that this quantitative evidence is due to physiological models that are capturing physiological characteristics that are relevant to interventions and intervention outcomes, but not captured by raw physiological variables. To investigate this hypothesis, we investigated whether the odds ratios associated with the latent variables were on par with those given to the raw features. In each of the tasks, latent state features were some of the most heavily weighted features for logistic regression (see Appendix). To identify which states are associated with high and low probabilities in weaning prediction, we then counted the frequency with which any particular model was associated with correctly predicting successful or unsuccessful weans. Specifically, we looked at which SSAM states generated the highest 1% of successful wean probabilities in the patients that were successful weans, and which SSAM states generated the lowest 1% of successful wean probabilities in the patients that were unsuccessful weans.

As shown in Figure 2-6, we see an increased membership in SSAM states 5 and 6 in those patients that had a high probability of a successful wean. On the other hand, data with a low probability of successful weaning in those patients who were not successful weans came more often from SSAM states 1 and 3. We then investigated the physiological variables that correspond to these states by examining the transitions probabilities for observed values in SSAM states 3 and 5 (recall that the state of the SSAM governs the dynamics of the observed physiological variables). There are several interesting differences in these probabilities. In SSAM state 5, transition probabilities for blood hematocrit values tended to stabilize from large abnormal values towards normalcy more often (8% vs. 5%). This could be indicative of patients who are healthy enough to remove fluid resuscitation, so their hematocrit is responding with decreased blood viscosity. In SSAM state 3, we observed that the respiration rate tends to stabilize from low values towards normalcy more often (13% vs. 11%). This could indicate that SSAM state 3 represents patients who eventually require some form of mechanical ventilation, which can cause more unsuccessful weaning patterns.

2.9 Discussion

Much literature in clinical prediction has focused on using large numbers of manually defined aggregate features as inputs to a classifier that will predict the risk of clinically significant events. [46, 48, 54] Switching dynamical systems models have been used to impute signals, identify artifacts, and discover physiological states in a variety of critical settings. [56, 75, 79] Most of these works have focused on developing models for densely sampled, often one-dimensional data. Our work differs in that we consider higher dimensional data and use discretization and binning to find relevant signals over longer time scales. Other work has applied unsupervised methods to discretized time series to discover anomalies and patient similarities, but without a latent variable representation. [88, 100] Time series symbolization creates many opportunities to analyze physiological data with the rich literature of techniques developed for discrete sequences; [59] our data processing approach also makes it natural for us to consider rich, nonlinear transition models, such as random forests, rather than the linear dynamical systems approaches of the work above.

The most recent prior work on vasopressor prediction used a subset of the MIMIC II patients receiving fluid resuscitation (2,944 adult ICU patients), and attempted to predict subsequent vasopressor administration within 2 hours using a general model and two disease-based models. [26] The general patient model achieved an AUC of 0.79 ± 0.02, and the disease-models had AUCs of 0.82 ± 0.02 for pneumonia and 0.83 ± 0.03 for pancreatitis. Our model used a similar short-term prediction approach in the general ICU population and achieved an AUC of 0.88 (± 0.0061). To our knowledge, we have the highest reported results for predicting vasopressor administration. These results suggest that the latent states discovered by the SSAM is an effective summary statistic for making predictions about the patientâĂŹs future intervention needs; an increased AUC of 0.05 could affect the treatment of thousands of patients annually in a large ICU.

Predicting weaning success is harder than predicting intervention onset. There exists fundamental uncertainty about when is the right time to wean a patient, and the decision may depend on staffing considerations, clinical judgment, or lack of familial support for intervention removal. In addition, unlike onset, the time of weaning is often present only in the patient note and not indicated in any structured data sources. The most relevant predictive work on vasopressor weaning specifically was done using clinically-guided feature engineering over sliding windows of data. [43] In particular, they selected 32 variables from a manually defined set over 438 clinically-guided features. They then classified patient segments that preceded successful vasopressor weaning by 1-12 hours (AUC = 0.81), and segments that preceded successful vasopressor weaning by 6-12 hours (AUC = 0.76). This was improved by only looking at those patients who survived their hospital admission to AUCs of 0.82 and of 0.825 respectively. While our AUCs are lower (0.71 ± 0.005), our approach did not use the large set of hand-engineered features; seeing whether our unsupervised physiological features improve prediction accuracy when combined with these engineered features is an interesting future direction. Another difference is that they excluded people who died in hospital, whereas we excluded people who died in a month after discharge.

We obtain AUCs of .092, 0,88, and 0.71 for predicting un-gapped vasopressor administration, gapped vasopressor administration, and vasopressor weaning. Our results for vasopressor use are the best achieved to our knowledge, and better results on vasopressor weaning were obtained with feature engineering on a smaller dataset. An important property of our approach is that our SSAM was trained in a completely unsupervised manner, specifically without knowing what the down-stream prediction task was to be, and without hand-specification of important features. Our goal in training the SSAM was to model the evolution of symbolized physiological time series—capturing global trends in the dynamics of the measurements that could be interpreted as the physiological state. The features derived from our SSAM resulted in improved performance regarding whether a patient would receive a vasopressor administration (our 0.88 AUC versus 0.79 AUC for gapped prediction; we also discovered several features associated with successful weaning from vasopressors and, to our knowledge, made the first attempt to quantify anecdotal claims about unnecessary intervention time. In summary, our work takes an important step toward moving away from hand-engineered, task-specific features to features that capture key information about patient health.

Our predictions of when a patient is ready to wean is just one of several actionable predictions in the space of vasopressor administration. Another important step would be to also consider the drug and dosage used for the vasopressor. In particular, a multicenter randomized trial comparing the use of dopamine or norepinephrine as first-line vasopressor therapy in 1,679 patients with shock found that patients treated with dopamine had significantly more arrhythmic events. [20] We could also improve the prediction quality of our model with additional features, such as those used to predict sepsis (sepsis is often preceded by episodes of hypotension, so an early predictor of sepsis could also be learning many of the states that might require vasopressor use). [39] Another interesting direction for future work would be to test whether these features assist in stratifying risk for a variety of interventions and intermediate outcomes, such as mechanical ventilation, [101, 110], sepsis [39], and response to different dosages of vasopressor [20] which, to date, have relied on hand-engineered features.

SSAMs have demonstrated value in detecting physiological states that influence the evolution of clinical measurements over time, and our overall methodology could be used to answer many other clinical questions. In the specific context of vasopressor weaning readiness, the ability to display the probability of a patient's possible need for an intervention, and their potential for weaning success, are important pieces of information that could enable clinicians to view predictions across entire ICU populations, updated on an hourly basis. This information could be further operationalized to create a clinical environment where potential therapies can be evaluated based on their prior performance in diverse populations and settings.

Feature Name	Intervention V-	Intervention V+	Control (C)
Age	65.812	65.811	60.787
% Male	65.719	65.719	56.070
SAPS-I	15.889	15.889	10.722
SOFA	7.844	7.844	3.251
Weight	82.229	82.229	81.767
ICU LoS	1.974	1.974	1.708
% Mortality	5.165	5.165	2.687
% MICU	14.251	14.251	41.733
% SICU	10.014	10.014	30.319
% CCU	12.759	12.759	16.336
% CSRU	62.085	62.085	9.504
% FICU	0.891	0.891	2.108
% Pacemaker Use	62.783	62.783	57.895
% ROF	59.726	59.725	5.660
Mean BP	76.235	74.680	82.120
TEMP	97.865	98.562	98.371
HR	83.979	85.251	83.682
SPO2	97.716	97.283	97.244
FIO2	0.736	0.530	0.51
RR	16.094	18.056	18.286
GLU	150.657	134.618	138.732
BICAR	25.090	24.023	24.866
HCT	29.031	29.903	31.457
K	4.531	4.244	4.078

Table 2.1: Mean values for variables in patient populations. ICU LoS denote ICU length of stay in days, ROF is Risk of Falls, BP is blood pressure, BMI is body mass index, HR is heart rate, SPO2 is the peripheral capillary oxygen saturation, FIO2 is the fraction of inspired oxygen, RR is the respiration rate, and HCT is hematocrit. Care units are medical (MICU), surgical (SICU), cardiac care (CCU), and cardiac-surgery recovery (CSRU).

Table 2.2: Performance of features in vasopressor need tasks using logistic regression classifier. Imminent Need predictions are inherently easier, as the data immediately prior to onset is available. Short-term Need predictions are more challenging because they enforce a time gap between observed data and the onset of the intervention. In general, SSAM features learned with Naive Bayes performed as well as the raw data, and the combination of SSAM features and Raw data did strictly better than either alone.

FEATURES USED	IMMINENT NEED PREDICTION (AUC)	SHORT-TERM NEED PREDICTION (AUC)
RAW	0.89 (± 1.1e-16)	0.83 (± 0.0040)
SSAM (RF)	0.81 (± 0.0584)	0.66 (± 0.0046)
SSAM (NB)	0.87 (± 0.0090)	0.83 (± 0.0076)
COMBINED: RAW+SSAM (RF)	0.92 (± 0.0008)	0.86 (± 0.0032)
COMBINED: RAW+SSAM (NB)	0.92 (± 0.0016)	0.88 (± 0.0061)

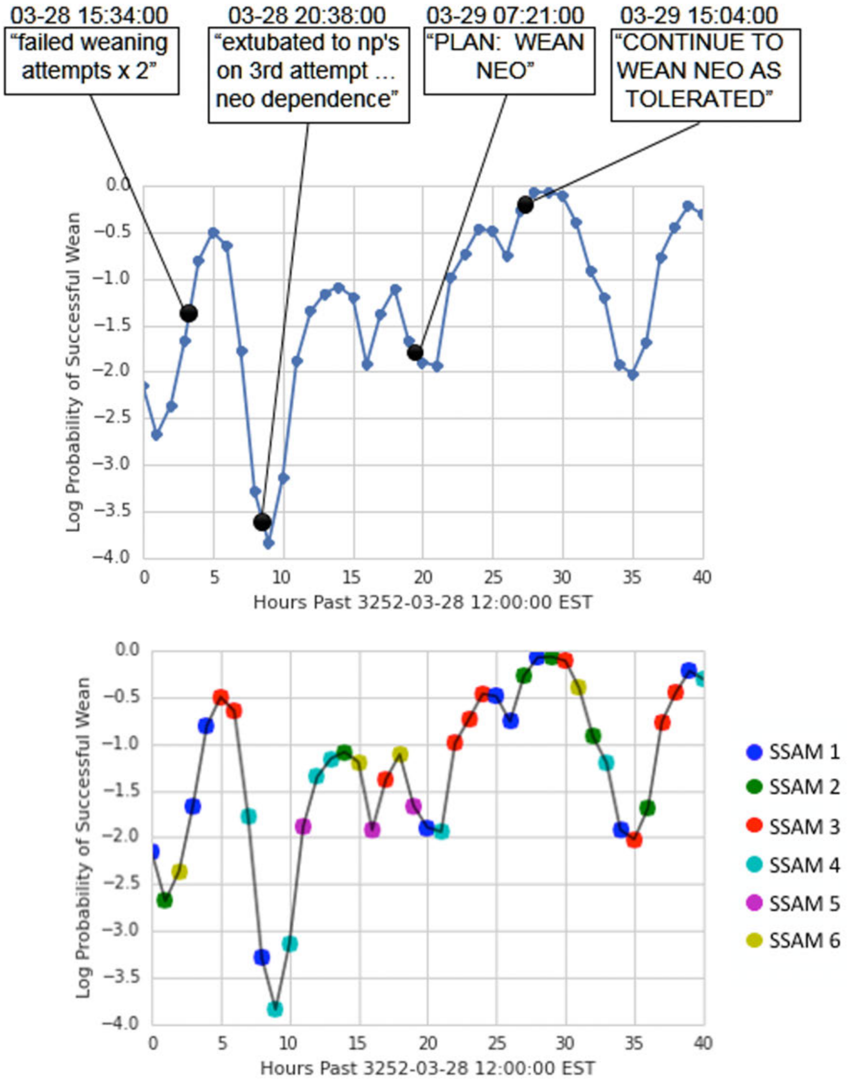
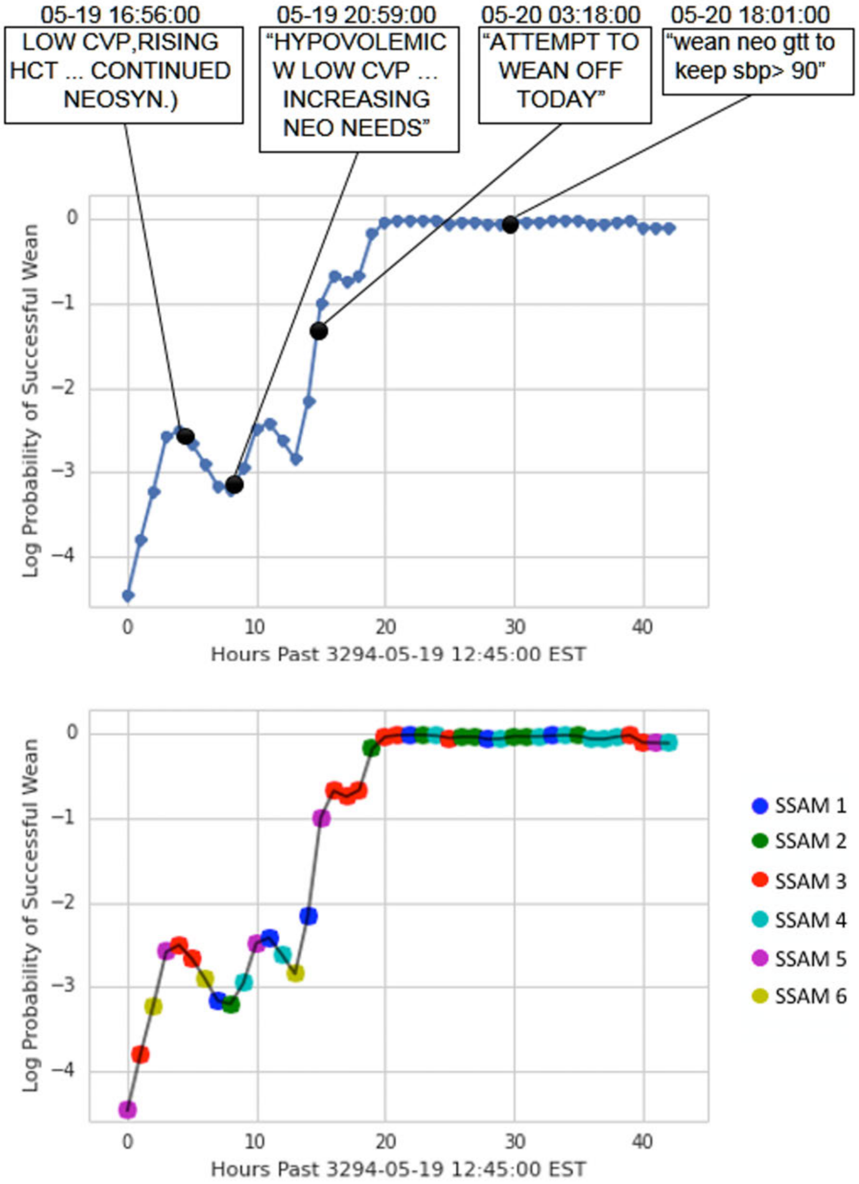


Figure 2-4: Probabilities of successful weaning and state for Case 1.



05-19 16:56:00
LOW CVP,RISING HCT ... CONTINUED NEOSYN.)

05-19 20:59:00
"HYPOVOLEMIC W LOW CVP ... INCREASING NEO NEEDS"

05-20 03:18:00
"ATTEMPT TO WEAN OFF TODAY"

05-20 18:01:00
"wean neo gtt to keep sbp> 90"

Figure 2-5: Probabilities of successful weaning and state for Case 2.

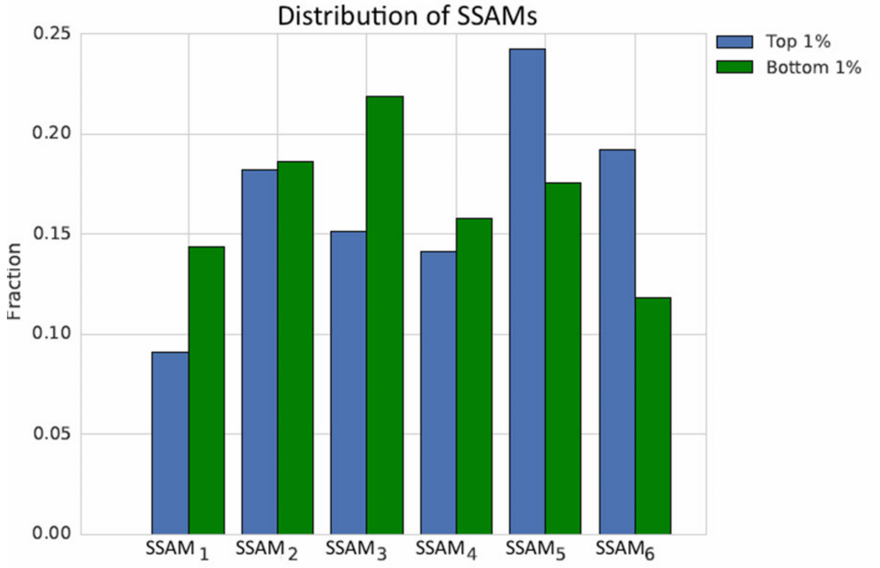


Figure 2-6: Histograms of the states across patients at time points of high (left) and low (right) probabilities of successful weans.

Chapter 3

Representations for Discharge and Post-Discharge Outcome Prediction

In this chapter, we focus on predicting discharge and post-discharge outcomes, including in-hospital mortality and post-dischrage mortality. In particular, we evaluate two representation techniques towards these tasks: topic modeling (or Latent Dirichlet Allocation) over the clinical notes, and Multi-Task Gaussian Processes over clinical timeseries (both phsyiological measurements and note/topic timseries). Most ICU mortality models primarily consider structured data [15, 43] or physiological waveforms [46, 88]. However, most do not consider the information captured in providers' free text notes, account for interventions given by care staff, or combine different forms of data. Many of the gold-standard ICU scores are also not intended to be continuous surrogates of patient status [52]. Early recognition of mortality could be used as a marker for physiological decline. The ICU is a location for critical decisions that weigh patient state against possible response to treatment.

3.1 Background

Modeling mortality in critical care settings has been a broad area of research. Siontis et al. [93] reviewed 94 studies with 240 assessments of 118 mortality prediction tools from 2009 alone. Many of these studies evaluated established clinical decision rules for predicting mortality, such as APACHE [47], SAPS-II [52], and SOFA [102] (with median reported AUCs of 0.77, 0.77, and 0.84, respectively). ICU scoring systems such as SAPS (simplified acute physiology score) use physiologic and other clinical data for acuity assessment. However, Sionitis et al. noted a large variability of these measures across various diseases and population subgroups. Even if the system itself were perfect, in 2012 the scoring systems were used in only 10% to 15% of US ICUs [8].

Work to score patient state has primarily focused on feature engineering for mortality prediction. This is usually accomplished by windowing or aggregating the structured numerical data so that a single feature matrix can be fed into a structured deterministic classifier. Some work has focused on clinical vitals and labs data for mortality and risk prediction. Hug et al. [43] used several hundred structured clinical variables to create a real-time ICU acuity score that reported an AUC of 0.88-0.89 for in-hospital mortality prediction. Weins et al. represented patient risk as a time series, and used the distance to the margin of an SVM for identifying patients at risk of testing positive for hospital acquired Clostridium difficile (AUC of 0.79) on a held-out set of several hundred patients. [109]

Some work has used the clinical text written by care staff as a fundamentally different type of signal: a trained professional's sparse representation of a patient's physiology over time. Topic models are latent variable models that incorporate information from free text notes to create topic-document-word mappings. [2,5] Several recent works have used information from clinical notes in their model formulations. Saria et al. [92] combined structured physiological data with concepts from the discharge summaries to achieve a patient outcome classification F1 score of 88.3. Similarly, [31] described preliminary results indicating that topic models extracted from clinical text in a subgroup of ICU patients were valuable in the prediction of per-admission mortality. They found that common topics from the unlabeled clinical notes were predictive of mortality, and an RBF SVM achieved a retrospective AUC of 0.855 for in-hospital mortality prediction using only learned topics. Lehman et al. [55] applied Hierarchical Dirichlet Processes to nursing notes from the first 24 hours for ICU patient risk stratification. They demonstrated that unstructured nursing notes were enriched with clinically meaningful information, and this information could be used for clinical support. Using topic proportions, the average AUC for hospital mortality prediction was 0.78 (± 0.01). In combination with the SAPS-I variable, their average AUC for hospital mortality prediction was 0.82 (± 0.003).

Other efforts have focused on combining well-engineered aggregate features with unsupervised clustering methods. In 2011, Kshetri et al. modeled patient states in intensive care patients using unsupervised clustering of multiple time-synched physiological signals. [48] Each cluster was examined post-hoc for concentrations of conditions of interest, including in-hospital mortality and acute kidney injury. The authors believe that the proportional concentration of certain states in different clusters (e.g. most patients who died passed through cluster 10) made for an interesting low-dimensional representation of patient state. Cohen et al. also used clustering analysis to identify clinically relevant patients states, but their analysis focused on physiological data obtained after patient trauma [18]. Joshi et al. infused more prior physiological knowledge into the clustering process using a layered technique known as Radial Domain Folding. [46] In this work, patient severity was modeled in the ICU by first transforming physiological signals into organ-system focused clusters. These clusters are then "folded" together to create aggregate estimates of the overall physiological state based on the permutations of cluster assignments at any given time. This approach requires that hand-engineered aggregate features be aggregated into a higher-level set of features for further prediction use.

3.2 Modeling Mortality Risk with Clinical Note Representations

3.2.1 Overview

We focus on the task of on-going mortality prediction in the ICU using clinical notes. The ICU is a particularly challenging environment because each patient's severity of illness is constantly evolving. Further, modern ICUs are equipped with many independent measurement devices that often produce conflicting (and even false) alarms, adversely affecting the quality of care. Consequently, much recent work in ICU mortality models [45, 52, 102] has aimed to consolidate data from these devices (primarily structured data and physiological waveforms) and transform these information streams into knowledge. However, these works omit perhaps the most descriptive sources of medical information: free-text clinical notes and reports.

The narrative in the clinical notes, recorded by expert care staff, is designed to provide trained professionals a quick glance into the most important aspects of a patient's physiology. Combining features extracted from these notations with standard physiological measurements results in a more complete representation of patients' physiological states, thus affording improved outcome prediction. Unfortunately, free-text data are often more difficult to include in predictive models because they lack the structure required by most machine learning methods. To overcome the obstacles inherent in clinical text, latent variable models such as topic models [2, 5] may be used to infer intermediary representations that can in turn be used as structured features for a prediction task.

We demonstrate the value of incorporating information from clinical notes, via latent topic features, in the task of in-hospital mortality prediction as well as 30 day and 1 year post-discharge mortality prediction. Specifically, we evaluated mortality prediction under three prediction regimes: (1) baseline regime, which used structured data available on admission (2) time-varying regime, which used baseline features together with dynamically accumulated clinical text using increasingly large subsets of the patient's narrative record, and (3) retrospective regime, which used all clinical text generated from a hospital stay to supplement the baseline features. In the time-varying regime, we also compare models based only on structured data to those also including topics from the notes. In all targeted outcomes, we demonstrate that adding information from clinical notes improves predictions of mortality.

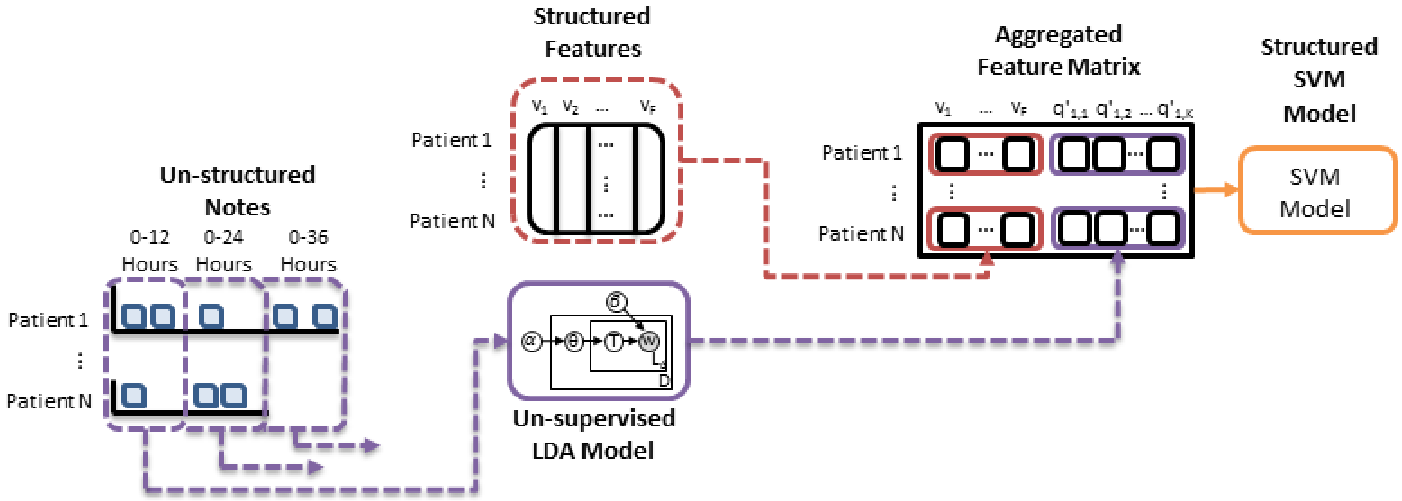


Figure 3-1: Overall flow of experiment. 1) Clinical baseline features are extracted from the database for every patient (e.g. age, sex, admitting SAPS-II score) and derived features are computed (e.g. maximum/minimum SAPS-II score) to form the *Structured Features* matrix v ($v_{p,f}$ is the value of feature f in the p^{th} patient). 2) Each patient's de-identified clinical notes are used as the observed data in an LDA topic model (i.e., *Un-supervisted LDA Model*), and a total of 50 topics are inferred to create the per-note topic proportion matrix q. 3) Per-note latent topic features are aggregated in extending 12 hour windows (e.g. notes within 0-12 hours, notes within 0-24 hours, etc.) and used to form matrix q' where $q'_{m,k}$ is the overall proportion of topic k in time-window m. 4) Depending on the model and time window being evaluated, subsets of the feature matrix v and matrix q' are combined into an *Aggregated Feature Matrix*. 5) A linear kernel SVM is trained to create classification boundaries for three clinical outcomes: in-hospital mortality, 30 day post-discharge mortality, and 1 year post-discharge mortality (i.e., *Structured SVM Model*).

3.2.2 Methods

Figure 3-1 gives a general overview of our experimental process. First, we extract clinical baseline features, including age, sex, and SAPS-II score, from the database for every patient. We also extract each patient's de-identified clinical notes. We use these notes as the observed data in an LDA topic model, and infer a total of 50 topics. We chose 50 topics after varying the number from $20 - 200$, and noting that validation set accuracy did not improve after 50. We normalize the word counts associated with each note, so that each note is represented by a 50-dimensional vector, summing to 1. These per-note topic distributions are then aggregated on a 12 hour semi-continuous timescale (e.g. notes within 0-12 hours, notes within 0-24 hours, etc.). A linear kernel SVM is trained to create classification boundaries with combinations of the structured clinical features and latent topic features to predict in-hospital mortality, 30 day post-discharge mortality, and 1 year post-discharge mortality.

3.2.3 Data and Pre-Processing

We used ICU data from the MIMIC II 2.6 database [87], a publicly-available, de-identified medical corpus that includes electronic medical records (EMRs) for $26,870$ ICU patients at the Beth Israel Deaconess Medical Center (BIDMC) collected from 2001 to 2008. Patient age, sex, SAPS-II scores, International Classification of Diseases-Ninth Revision (ICD-9) diagnoses, and Disease-Related Group were extracted. Medical co-morbidities were represented by the Elixhauser scores (EH) for 30 co-morbidities as calculated from the ICD-9 codes. Patient mortality outcomes were also queried to determine which patients died in-hospital, died within a certain time after discharge, or lived past the most recent query of Social Security records.

We extracted all clinical notes recorded prior to the patient's first discharge, including notes from nurses, physicians, labs, and radiology. The discharge summaries themselves were excluded because they typically stated the patient's outcome explicitly. Vocabularies for each note were generated by first tokenizing the free text and then removing stopwords using the Onix stopword list [1]. A TF-IDF metric [90] was applied to determine the 500 most informative words in each patient's notes, and we then limited our overall vocabulary to the union of the most informative words per-patient. This pre-processing step reduced the overall vocabulary down to 285,840 words from over 1 million terms while maintaining the most distinctive features of each patient.[2]

Patients were excluded if their notes had fewer than 100 non-stop words or were under the age of 18. Specific notes were excluded if they occurred after the the end of the day

[0] Note that MIMIC supports ICD-9-CM codes, which are the U.S. "clinical modifications" that support the use of the codes in billing.

[1] Onix Text Retrieval Toolkit, API Reference, http://www.lextek.com/manuals/onix

[2] Some medical term canonicalization parsers were also examined, but we found their outputs to be fairly unreliable for this task.

Table 3.1: Clinical Note Cohort Composition

	Train	Test	Total
Patients	13,524	5,784	19,308
Notes	331,635	142,129	473,764

in which a patient died or was discharged (e.g. radiology or lab reports whose results were reported afterwards). The resulting cohort consisted of 19,308 patients with 473,764 notes. We held out a random 30% of the patients as a test set. The remaining 70% of patients were used to train our topic models and mortality predictors. Table 3.1 summarizes the number of notes and patients in the training and test sets.

3.2.4 Structured and Derived Features

In total, we extracted and derived 36 structured clinical variables for each patient: the age, gender, SAPS II score on admission, minimum SAPS II score, maximum SAPS II score, final SAPS II score, and the 30 EH comorbidities. Data were scaled to avoid the range of a feature impacting its classification importance. This formed a feature matrix v, where the element $v_{p,f}$ was the value of feature f in the p^{th} patient.

3.2.5 Topic Inference

Instead of considering each note separately, we used the set of all notes that occurred in a particular time period as features for that period. We examined the distribution of note times, and found three peaks in note entry for any given day in a patient's stay (e.g. day 1, day 2, etc.): around 06:00, 18:00 and 24:00.[3] Given this distribution, we used 12 hours for our time windows.

Topics were generated for each note using Latent Dirichlet Allocation [5,37]. Our initial experiments found no significant difference in held-out prediction accuracy across a range of 20 to 100 topics. We set hyperparameters on the Dirichlet priors for the topic distributions

[3] The increases in note submission at 06:00 and 18:00 were likely due to the current 12 hour nursing shift cycle. The large number of notes submitted at end-of-day were likely due to a previously common 14:00 - midnight nursing shift.

(α) and the topic-word distributions (β) as $\alpha = \frac{50}{numberTopics}, \beta = \frac{200}{numberWordsInVocab}$. We used 50 topics in our final experiments, and topic distributions were sampled from an MCMC chain after 2,500 iterations. This topic-modeling step resulted in a 50-dimensional vector of topic proportions for each patient for each note.

We concatenated the topic vectors into a matrix q where the element $q_{n,k}$ was the proportion of topic k in the n^{th} note. Of particular interest was whether certain topics were enriched for in-hospital mortality and long-term survival. We used an enrichment measure defined by Marlin et al. [61], where the probability of mortality for each topic is calculated as $\theta_k = \frac{\sum_{n=1}^{N} q_{n,k} * y_k}{\sum_{n=1}^{N} q_{n,k}}$, where y_n is the noted mortality outcome (0 for a patient who lives, and 1 for a patient who dies). These enrichment measures are reported in section 3.2.7.

The time windows were used to construct feature vectors for each prediction task, where (at each step) we extended the period of consideration forward by 12 hours. From the previously constructed per-note matrix q that describes the distribution over topics in each note, we collapse into another matrix q' where $q'_{m,k}$ describes the overall proportion of topic k in time-window m. The element $q'_{m,k}$ is given by the mean of that topic's proportions of all the notes in time-window m: $\text{mean}_{n \in m} q_{n,k}$.

3.2.6 Prediction Task Definition

We considered three prediction regimes with the inferred topic distributions: baseline prediction, dynamic (time-varying) outcome prediction and retrospective outcome prediction for the outcomes of in-hospital, 30-day, and 1-year mortality.

A separate linear SVM [12] was trained for each of the three outcomes, and each set of model features evaluated. The loss and class weight parameters for the SVM were selected using five-fold cross-validation on the training data to determine the optimal values with AUC as an objective. The learned parameters were then used to construct a model for the entire training set, and make predictions on the test data.

All outcomes had large class-imbalance (mortality rates of 10.9% in-hospital, 3.7% 30 day post-discharge, and 13.7% 1 year post-discharge[4]). To address this issue, we randomly sub-sampled the negative class in the training set to produce a minimum 70%/30% ratio between the negative and positive classes. Test set distributions were not modified to reflect the reality of class imbalance during prediction, and reported performance reflects those distributions.

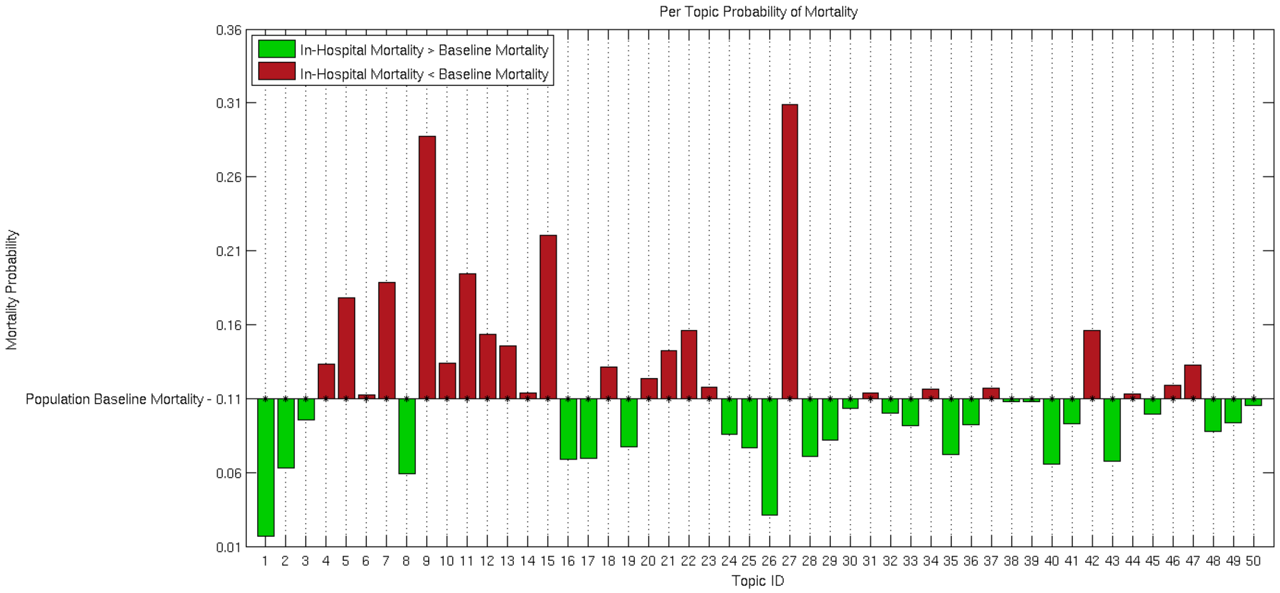


Figure 3-2: The probability of in-hospital mortality for each topic, indicating that topics represent differences in outcome. Probabilities are calculated as $\theta_k = \frac{\sum_{n=1}^{N} q_{n,k} * y_n}{\sum_{n=1}^{N} q_{n,k}}$ (see section 3.2.5). Each bar shows the prevalence of a given topic k in the mortality category, as compared to the set of all patients. Bars are shown as above (in red) or below (in green) the baseline in-hospital mortality based on the value of θ_k for each topic k.

First, we established a static baseline model using only structured features present at admission (i.e. clinical baseline features and derived features thereof). We then ran dynamic outcome prediction in intervals of 12 hours at each step by including larger sets of patient notes in a step-wise manner. We finally performed retrospective outcome predictions, where we included structured features and all notes written during the stay as a static entity for prediction. Significantly, predictions of mortality with this type of feature set are a

[4] This includes those who die within the first 30-days post-discharge, so two of the prediction targets have overlap.

[4] Note that we purposefully excluded the discharge summaries

retrospective exercise only: it is not possible to first select all notes that occur before a patient's death, and then predict in-hospital mortality, because the time of mortality is not known a-priori. The observer would have to "know" that the patient's hospital record was about to finish (either by death or discharge). The following settings were evaluated:

- *Admission Baseline Model*: A baseline model using the structured features of age, gender, and the SAPS II score at admission. These baseline features are extracted from the data present at patient admission only. (3 features total)

- *Time-varying Topic Model 1 - 20*: Outcome prediction performed by including notes in a step-wise fashion, extending the period of consideration forward by 12 hours at each step. For example, Time-varying Topic Model 1 includes topic features derived from all notes written during the first 12 hours of a patient's stay in the ICU, while Time-varying Topic Model 20 includes those derived from the first 240 hours. (50 features total)

- *Combined Time-varying Model 1 - 20*: Outcome prediction using the same setup as Time-varying Topic Model 1 - 20, but with the static structured features from Admission Baseline Model (gender, age, admitting SAPS score) included. (53 features total)

- *Retrospective Derived Features Model*: A retrospective model using the structured features of age, gender, admitting SAPS II score, the minimum SAPS II score, the maximum SAPS II score, the final SAPS II score, and all EH comorbidities. (36 features total)

- *Retrospective Topic Model*: A retrospective model using topics derived from all notes written during a patient's stay in the ICU. (50 features total)

- *Retrospective Topic + Admission Model*: A retrospective model combining structured features from Admission Baseline Model (gender, age, admitting SAPS scores) with latent topic features from Retrospective Topic Model. (53 features total)

- *Retrospective Topic + Derived Features Model*: A retrospective model combining structured features from Retrospective Derived Features Model (gender, age, admitting/min/-max/final SAPS scores, EH comorbidities) with latent topic features from Retrospective Topic Model. (86 features total)

We compare the prediction results for all models on each of the outcomes in Figure 3-3 and Table 3.3. We again emphasize that retrospective models are retrospective exercises only to establish the isolated and combined prediction ability of clinical notes and features. We also note that our *Time-varying Topic Model* is time-varying only in its application. We do not use other possible latent variable models such as "Dynamic topic models" [6], because we do not want to model the time evolution of topics, but rather the time evolution of membership to a given set of topics.

3.2.7 Results

Qualitative Enrichment Table 3.2 lists the top words for the topics that had the largest enrichment ($\theta_k = \frac{\sum_{n=1}^{N} q_{n,k} * y_k}{\sum_{n=1}^{N} q_{n,k}}$) for in-hospital mortality, the smallest enrichment for in-hospital mortality, and the highest enrichment for 1 year mortality. The relative distributions of the in-hospital mortality probabilities for each of the 50 topics are shown in Figure 3-2. There was a wide variation in the in-hospital mortality concentration for the different topics, ranging from 3% - 30%. (See Table A.2 for a listing of top ten words for all topics.)

The topics enriched for in-hospital mortality presented a detailed view of the possible causes of death in the ICU. For example, patients in a modern ICU rarely die suddenly. Often patient life is sustained for some time in order for their family to express their wishes regarding terminal care and death. This could be one interpretation for Topic 27, which pertains to the discussion of end-of-life care options. Other topics with in-hospital mortality enrichment pertained to top causes of ICU mortality: respiratory infection (Topic 7), respiratory failure (Topic 15), and renal failure (Topic 5).

Hospital survival was also marked by topics which seem relevant to factors tied closely to the ability to recover from physiological insults: patients who are admitted for cardiovascular

Table 3.2: Top ten words in topics enriched for in-hospital mortality, hospital survival (any number of days post-discharge), and 1 year post-discharge mortality.

	Topic	**Top Ten Words**	**Possible Topic**
In-hospital Mortality	27	name, family, neuro, care, noted, status, plan, stitle, dr, remains	Discussion of end-of-life care
	15	intubated, vent, ett, secretions, propofol, abg, respiratory, resp, care, sedated	Respiratory failure
	7	thick, secretions, vent, trach, resp, tf, tube, coarse, cont, suctioned	Respiratory infection
	5	liver, renal, hepatic, ascites, dialysis, failure, flow, transplant, portal, ultrasound	Renal Failure
Hospital Survival	1	cabg, pain, ct, artery, coronary, valve, post, wires, chest, sp	Cardio-vascular surgery
	40	left, fracture, ap, views, reason, clip, hip, distal, lat, report	Fracture
	16	gtt, insulin, bs, lasix, endo, monitor, mg, am, plan, iv	Chronic diabetes
1 Year Mortality	3	picc, line, name, procedure, catheter, vein, tip, placement, clip, access	PICC[5] line insertion
	4	biliary, mass, duct, metastatic, bile, cancer, left, ca, tumor, clip	Cancer treatment
	45	catheter, name, procedure, contrast, wire, french, placed, needle, advanced, clip	Coronary catheterization

surgery (Topic 1) are often not allowed as surgical candidates until they are in relatively good health; patients who are able to respond to their care staff and the ICU environment (Topic 26, Table A.2) are adequately dealing with the known stress of ICU admission; patients with trauma-based injuries such as fracture and pneumothorax (Topics 8, 40); and patients with chronic conditions like diabetes (Topic 16).

The topics enriched for 1 year post-discharge mortality suggested that patients who are discharged but die within a year have conditions with a low chance of long-term survival. For example, cancer (Topic 4), the need for long-term IV access while in the ICU (Topic 3), and the use of coronary catheterization (Topic 45) to diagnose activity in coronary arteries or other valvular/cardiac issues.

Prediction We evaluated the predictive power of each model and outcome pair. Figure 3-3 shows the AUCs achieved by each model for the three targeted outcomes. Table 3.3 lists a more complete set of the SVM classification metrics.

As shown in Table 3.3, the prevalent class imbalance resulted in a bias toward low specificities in the *Admission Baseline Model*. The balance between sensitivity and specificity generally leaned towards favoring higher specificities for in-hospital and 30 day mortality prediction as time moved forward in the *Time-varying* models, but this was not uniformly true in all cases. In general, the *Retrospective Derived Features Model* had a high sensitivity and low specificity, the *Retrospective Topic Model* had good specificity, and the combined models tended to have a more even set of both measures.

For 30 day and 1 year post-discharge mortality prediction, the *Admission Baseline Model* was very steady, averaging an AUC of 0.68 over all time windows for both outcomes. The *Combined Time-varying Model* achieved an average/best performance of 0.77/0.8 for 30 day mortality and 0.75/0.77 for 1 year mortality. In both outcomes the *Time-varying Topic Model* performed strictly better than the *Admission Baseline Model* until the available patient subset became minimal (the 204 -216 hour windows), and the *Combined Time-varying Model* was always better than either alone.

As expected, the four *Retrospective* models were generally more predictive than any of the *Time-varying* models. *Retrospective* models tended to increase performance as more features were added. For in-hospital and 30 day mortality prediction, the *Retrospective Topic Model* performed better than the *Retrospective Derived Features Model* (AUCs increased from 0.90 to 0.94 and 0.75 to 0.78 respectively). For 1 year mortality this was reversed (AUC decreased from 0.78 to 0.76).

In the in-hospital mortality setting, it seemed that admission features were not needed once latent topic features are known, but the derived features did provide extra information[6]. However, in the 30 day setting, latent topic features were similarly improved by either

[6] Adding the admission features did not improve the *Retrospective Topic Model*, but adding the derived features boosted

the admission features or the derived features[7]. This is likely because the derived features included EH comorbidities derived from the ICD-9 codes, and the ICD-9 codes themselves are often transcribed after a patient's discharge with the most actionable (or billable) conditions a patient presented. It is possible that these features are most relevant to in-hospital mortality risks (e.g. EH scores for myocardial infarction, congestive heart failures, etc.). However, this also suggests that the EH scores are not a practical way to build predictive models, because they cannot be computed until after discharge.

3.3 Modeling Clinical Timeseries with Gaussian Process Hyperparameter Representations

3.3.1 Overview

The general issue of comparing signals that are not aligned and irregularly sampled has been considered before. Establishing similarity metrics among time-series data is an important part of many learning tasks and often is achieved using a variety of summarization methods. However, many modeling methods fail when applied to irregularly sampled data unless strong assumptions are made about the functional form present in the underlying data source. Furthermore, in cases where such methods work, data imputation is often necessary, which can introduce additional sources of error and bias. Finally, many methods work on a single timeseries, but fail to generalize to (or take advantage of) other related time-series data.

This work was done in close collaboration with Marco Pimentel of Oxford University. Our proposed technique transforms a variety of irregularly-sampled clinical data into a new latent space using the hyperparameters of multi-task Gaussian Process (MTGP) models. Patients are compared based on their similarity in the new hyperparameter space. Our work differs from other work in that it: 1) uses the correlation between and within multiple time-series to estimate parameters instead of considering each timeseries separately; 2) infers a compact

AUC slightly to 0.96.

[7] Adding the admission features to the *Retrospective Topic Model* improved AUC to 0.81 but adding the derived features did not improve AUC further.

latent representation of the source data, rather than finding patterns that are common within different timeseries; and 3) leverages the information contained in the inferred model hyperparameters for supervised learning, whereas others use the predicted mean function of the GP as a pre-processing or smoothing step (see 3.3.2).

We use MTGPs for forecasting patient acuity based on irregularly sampled heterogeneous clinical data. We evaluate the value of the inferred MTGP hyperparameters as a new latent space for representing multi-dimensional timeseries in two ways: 1) estimating and forecasting a cerebrovascular autoregulation index from noisy physiological time-series data in patients who suffered a traumatic brain injury and 2) transforming irregular ICU patient clinical notes into timeseries, and using MTGP hyperparameters from these timeseries as features to predict mortality probability.

3.3.2 Gaussian Processes

Gaussian processes (GP) form the basis for a Bayesian modeling technique that has been used for various machine learning tasks [83]. Most commonly, GPs are used to predict a single output (denoted here as a "task") based on one or more input timeseries. We refer to this model as a single-task GP (STGP). Lasko et al. used Gaussian process regression as a smoothing function of irregularly-sampled signals [50]. This is a common usage model for GPs on clinical timeseries: GPs are used to model observed data through the predicted mean function of the timeseries. Clifton et al. used GPs as a framework for coping with data artifacts and incompleteness in mobile sensor data [17]. In a related work [16], a functional version of extreme value statistics was proposed for physiological data in order to compare different timeseries. Similarly, GPs were used for robust regression of noisy heart rate data [95]. The remainder of the related work has used STGP models to predict a single output based on one or more input variables.

In the present study, we explore the potential of a novel approach using MTGP models [7] to learn the correlation between and within time-series, and obtain a concise representation of time-varying physiological and clinical data based on the inferred hyperparameters.

Here, we motivate the use of MTGPs and describe the method (source code is available on-line[8]) that we have adapted for hyperparameter construction [22].

3.3.3 Multi-Task Gaussian Process Models

The general STGP framework may be extended to the problem of modeling m tasks simultaneously where each model uses the same index set $\mathbf{x}$ (e.g., physiological or clinical timeseries). A naïve approach is to train a STGP model independently for each task, as illustrated in Figure 3-4(a). We introduce instead an extension to multi-task GP models proposed in [7], which makes use of the covariance in related tasks to reduce uncertainty in the inferred signal.

Let $\mathbf{X_n} = \{x_i^j \mid j = 1, ..., m, i = 1, ..., n_j\}$ and $\mathbf{Y_n} = \{y_i^j \mid j = 1, ..., m, i = 1, ..., n_j, \}$ be the training indices and observations for the m tasks, where task j has n_j number of training data. We consider the regression model $\mathbf{y}_n = g(\mathbf{x}_n) + \epsilon$, in which $g(x)$ represents the latent function and $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$ is a noise term. GP models assume that the function $g(\mathbf{x}_n)$ can be interpreted as a probability distribution over functions such that $\mathbf{y_n} = g(\mathbf{x}_n) \sim \mathcal{GP}\Big(m(\mathbf{x}_n), k(\mathbf{x}_n, \mathbf{x}_n')\Big)$, where $m(\mathbf{x}_n)$ is the mean function of the process (assumed $= 0$) and $k(\mathbf{x}_n, \mathbf{x}_n')$ is a covariance function describing the coupling among the independent variables $\mathbf{x}_n$ as a function of their kernel distance. To specify the affiliation of index x_i^j and observation y_i^j to task j, a label $l^j = j$ is added as an additional input to the model, as shown in Figure 3-4(b). To model the correlation between tasks as well as the temporal behaviour of the tasks within a unified GP model, two independent covariance functions are assumed, and the covariance matrix $\mathbf{K}_{MT}$ for all m tasks can be written

$$\mathbf{K}_{MT}(\mathbf{X}_n, \mathbf{l}, \boldsymbol{\theta}_c, \boldsymbol{\theta}_t) = \mathbf{K}_c(\mathbf{l}, \boldsymbol{\theta}_c) \otimes \mathbf{K}_t(\mathbf{X}_n, \boldsymbol{\theta}_t) \tag{3.1}$$

where $\otimes$ is the Kronecker product, $\mathbf{l} = \{j \mid j = 1, ..., m\}$, $\mathbf{K}_c$ and $\mathbf{K}_t$ represent the correlation and temporal covariance functions, and $\boldsymbol{\theta}_c$ and $\boldsymbol{\theta}_t$ are vectors containing hyperparameters for $\mathbf{K}_c$ and $\mathbf{K}_t$, respectively. Within geostatistics, this approach is also known as the *intrinsic*

[8] `http://www.robots.ox.ac.uk/~davidc/publications_MTGP.php`

correlation model [106].

By modifying the temporal covariance function we can encode our prior knowledge concerning the functional behavior of the tasks that we wish to model. The most frequently-used example is the squared-exponential covariance function [83]:

$$\mathbf{K}_t = \theta_A^2 \exp\left\{ -\frac{\parallel x - x' \parallel^2}{2\theta_L^2} \right\}, \tag{3.2}$$

where $\boldsymbol{\theta}_t = \{\theta_A, \theta_L\}$, and θ_A and θ_L are hyperparameters modeling the y-scaling and x-scaling (or time-scale if the data are timeseries) of the covariance function, respectively.

To construct a valid positive semidefinite correlation covariance function $\mathbf{K}_c$, we used the Cholesky decomposition and the "free-form" parameterization of the elements of the lower triangular matrix $\mathbf{L}$ proposed in [7], such as

$$\mathbf{K}_c = \mathbf{L}\mathbf{L}^\top, \quad \mathbf{L} = \begin{bmatrix} \theta_{c,1} & 0 & \dots & 0 \\ \theta_{c,2} & \theta_{c,3} & & 0 \\ \vdots & & \ddots & \vdots \\ \theta_{c,k-m+2} & \theta_{c,k-m+2} & \dots & \theta_{c,k} \end{bmatrix} \tag{3.3}$$

where $k = \sum_{i=1}^{m} i$ is the number of correlation hyperparameters.

Identically to STGPs, the hyperparameters $\boldsymbol{\theta}$ for a MTGP may be optimized by minimizing the negative log marginal likelihood via gradient descent [83], and predictions for test indices $\{\mathbf{x}_p^*, \mathbf{l}_p^*\}$ can be made by computing the conditional probability $p(\mathbf{y}_p^* | \mathbf{x}_p^*, \mathbf{l}_p, \mathbf{x}_n, \mathbf{l}_n, \mathbf{y}_n)$.

Figure 3-5 shows an example of STGPs and an MTGP applied to a simple synthetic dataset with 4 sample tasks. Tasks 1 and 2 were correlated, task 1 and task 2 were both anti-correlated with task 4, and task 3 was uncorrelated with all other tasks. For this, 4 tasks were sampled from a MTGP model with the following hyperparameters: $\theta_L = \theta_A = \theta_{c,1} = \theta_{c,2} = \theta_{c,3} = \theta_{c,6} = \theta_{c,10} = 1$, $\theta_{c,4} = \theta_{c,5} = \theta_{c,0} = 0$, and $\theta_{c,7} = \theta_{c,8} = -1$. Artificial gaps were then randomly created in different tasks at different time points and with different durations. The STGP (Figure 3-5(b)), applied to each task independently, fails to adequately represent the functions, particularly where data are not available. Figure 3-5(c) shows that the MTGP improves the predictions in all 4 tasks by capturing the relationships between them.

The MTGP has several useful properties as compared to the traditional GP:

- We can allow task-specific training indices n_j; i.e., training data may be observed at different times for different tasks (Figure 3-5);
- The correlations within and between tasks are automatically learned from the data by fitting the covariance function in Equation 3.1; and
- The framework assumes that the tasks have similar temporal characteristics and hyperparameters θ_t.

A limitation of the MTGP is computational cost: $\mathcal{O}(m^3n^3)$ compared with $m \times \mathcal{O}(n^3)$ for STGPs. This limitation is not as relevant for our application, given that we are not dealing with densely-sampled time-series data, but data which is sparse and irregular. Another limitation of the MTGP is that the number of hyperparameters can increase rapidly for an increasing number of tasks, which can lead to a multi-modal parameter space.

3.3.4 Signal Representation via Hyperparameters

We use the inferred MTGP hyperparameters $\boldsymbol{\theta}$ that describe the temporal correlation within and between tasks as features that represent our set of observations: θ_A and θ_L which respectively govern each output scale of our functions and the input, or time, scale, and $\theta_{c,i}$ that correspond to the correlation between the different tasks (outputs) modelled. In effect, $\boldsymbol{\theta}$ provides a new latent search space to examine and evaluate the similarity of any two given multi-dimensional functions. Importantly, these parameters are:

1. a means of representing the functional behavior of a set of observations $\{\mathbf{y}_n, \mathbf{x}_n\}$;
2. learned directly from data; and
3. generalizable to any type of longitudinal data, including categorical and numerical types.

3.3.5 Experiment 1: Using Multiple Noisy Time-Series Data to Interpolate Clinical Signals and Assess Stability

In this experiment, we use physiological signals from Traumatic Brain Injury (TBI) patients to test the MTGP's ability to assess and forecast multiple related signals. We examine two noisy timeseries: the intracranial pressure (ICP) and mean arterial blood pressure (ABP). Continuous monitoring of ICP and ABP has become a standard in neurological ICUs. Cerebrovascular autoregulation is an important mechanism to sustain adequate cerebral blood flow [108], and impairment of this mechanism indicates an increased risk to secondary brain damage and mortality [41].

Cerebrovascular autoregulation is most commonly assessed based on the Pressure-Reactivity Index (PRx), which is defined as a sliding window Pearson's correlation between the ICP and ABP [19]. However, the ICP and ABP timeseries are often contaminated by artifacts and missing data, and PRx can no longer be calculated in these situations. Although methods have been proposed to detect and remove artifacts [25], the artifact removal process still creates gaps of missing data in the timeseries.

In this experiment, we demonstrate how the proposed MTGP model can be applied to interpolate the incomplete data in ICP and ABP signals and, more importantly, to accurately estimate PRx.

The ICP and ABP data were collected from 35 TBI patients who were monitored for more than 24-hours in a Neuro-ICU of a tertiary care hospital between January 2009 and December 2010. The continuously monitored physiological readings were sampled and recorded every 10 seconds. For experimental evaluation, we selected 30 ten-minute windows from each patient recording, where ICP and ABP signals were free from artifacts and missing values. We then randomly introduced artificial gaps in both signals as shown in Figure 3-6. We evaluated the PRx estimation accuracy, and we further compared the performance of MTGP to that of STGP, which models each signal independently. For implementation, priors over the hyperparameters were selected after 100 random initializations for each case.

Results The quality of predictions is evaluated using the squared error loss, where we compute the squared residual $(y^* - \hat{y}^*)^2$ between the mean prediction $(\hat{y}^*)$ and the target (y^*) at each test point, and the square root of the average over the test set to produce the root mean squared error (RMSE). As the RMSE is sensitive to the overall scale of the target values, we additionally evaluate the negative log probability of the target under the model, by defining the mean standardized log loss (MSLL) as

$$\begin{aligned}\text{MSLL}(\hat{\mathbf{y}}^*, \mathbf{y}^*) = \frac{1}{p} \sum_{i=1}^{p} \Big(& - \log\, p(\hat{y}_i^* | f, x_i^*) \\ & + \log\, p(\hat{y}_i^* | m(\mathbf{y}_n), \text{var}(\mathbf{y}_n), x_i^*\Big),\end{aligned}$$

where the first term is the log likelihood of $\hat{y}_i^*$ given our latent function f and the test index x_i^*. This probability is normalized by the second term, the log likelihood of $\hat{y}_i^*$ under a trivial model that predicts using a Gaussian with mean $m(\mathbf{y}_n)$ and variance $\text{var}(\mathbf{y}_n)$ of the training labels.

Table 3.4 shows the overall performance of our approach. We note that the MTGP was able to estimate the correlation between the ICP and ABP signals—PRx—accurately even with incomplete data. The average RMSE between the true correlation coefficients and the MTGP estimated ones with the incomplete data was 0.09 (Table 3.4). This suggests that the posterior hyperparameter of MTGP, which measures the interactions between ICP and ABP, may be used as an index to model the cerebrovascular autoregulation mechanism and thus the risk of secondary brain injury.

We note that the scale of ICP values is normally between 1 to 20 mmHg, and the specific ICP value determines whether the achieved reduction in RMSE is clinically significant. If the ICP has already elevated to somewhere near 20 mmHg, any slight increase in ICP may result in secondary damage to the brain. In this case, even small reductions to RMSE are desirable to guide the medical interventions.

We also observe that the MTGP provides a significant improvement in interpolating values for both signals, as the correlation between the two physiological variables is taken into account. Particularly, in periods of incomplete data (see Figure 3-6), the predictions are much more accurate compared to STGP. This shows that the proposed MTGP model can also be used for accurate interpolation and forecasting of ICP and ABP timeseries in the applications of advanced alarming and physiological trajectory analysis.

3.3.6 Experiment 2: Using Clinical Notes as Timeseries for ICU Mortality Prediction

To demonstrate the effectiveness of the proposed MTGP model on features inferred from sparse, irregularly sampled timeseries, we applied MTGPs to clinical notes from the ICU for mortality prediction as summarized in Figure 3-7.

Similarly to the data used in 3.2.3, we used 2001–2006 ICU data from the open-access MIMIC II 2.6 database [89]. For each patient we extracted the SAPS-I score, calculated from clinical variables over a patient's first 24-hours in the ICU. We used all notes from nursing, physicians, labs, and radiology recorded prior to the patient's first discharge from ICU. Discharge summaries were excluded because they typically state the patient's outcome explicitly. Patients were excluded if their notes had fewer than 100 words, fewer than 6 total notes in their record, or were under the age of 18. Patient mortality outcomes were measured at hospital discharge and 1 year post-discharge.

The final cohort consisted of 10,202 patients with 313,461 notes. A random 30% of the patients (3,040) were held back as a test set. The remaining 70% of patients (7,162) were used to train topic models and mortality predictors. The test set contained 93,411 notes, and the training set had 220,005.

Beginning from sparse, irregularly sampled clinical notes, we first performed topic modeling as a form of dimensionality reduction as described in section 3.2.5. The topic inference resulted in a 50-dimensional vector of topic proportions for *each note* in every patient's record. We concatenated topic vectors into a matrix q where the element q_{nk} was the proportion of topic k in the n^{th} note.

Hyperparameter Construction Once notes were transformed into multi-dimensional numeric vectors, we used the MTGPs to model the per-note change in topic membership over a patient's stay. This is critical for comparing two patients' records given that patients have different lengths of stay and note taking intervals depend on staff, clinical condition, and other factors.

From the topic enrichment measure (ϕ), we chose the topics with a posterior likelihood above or below 5% of the population baseline likelihood across topics. This yielded nine topics (see Table **??** for a summary of the chosen topics). We employed MTGP to learn the temporal correlation between the nine topics and the overall temporal variability of the multiple timeseries.

From the available data sources, we formed a set of three feature matrices: (1) the admitting SAPS-I score for every patient, (2) the average topic membership for the nine identified topics (matrix q), and (3) the inferred MTGP hyperparameters across the nine topic vectors from q. Importantly, the admitting SAPS-I score and mean topic members (1 and 2) are both *static measures*. SAPS-I collapses data from the first 24 hours of the record, while the average topic membership collapses the entire per-note timeseries for each patient's record into an aggregate measure. Our proposed MTGP hyperparameters (3) complement these measures with information about the per-note timeseries.

Outcome Classification We considered five feature prediction regimes that combined subsets of the feature matrices 1, 2, and 3 as an aggregate feature matrix. We trained two supervised classifiers that were identical in the five feature sets used, but provided different objective functions for optimization: Lasso logistic regression and L2 linear kernel SVM.

Classifiers were trained to create classification boundaries for two clinical outcomes: in-hospital mortality and 1-year post-discharge mortality. All outcomes had large class-imbalance (e.g., in-hospital mortality rates of 10.9%). To address this issue, we randomly sub-sampled the negative class in the training set to produce a minimum 70%/30% ratio between the negative and positive classes. Test set distributions were not modified, and reported performance reflects those distributions. Due to space constraints, we only report results on a completely held out test set. We performed 5-fold cross-validation on the remaining data, and cross-validation results were similar to those obtained on the completely held-out test set.

We evaluated the performance of all classifiers using the area under the Receiver Operating Characteristic curve (AUC) on the held-out test set. Table 3.6 reports results from the Lasso model. Results obtained using the L2 linear kernel SVM were not statistically different.

Results SAPS-I had the poorest predictive power, which is understandable given that it is only an initial snapshot (24 hours) of the severity of illness. We used the static SAPS-I score due to its status as the gold-standard in clinical scoring, and our argument in the second experiment is that the MTGP hyper-parameter space complements this clinical score, rather than competes with it. The average value of the most significant topics significantly improved upon that predictive power. The performance of MTGP Hyperparameters on their own was similar to that of the Topics: AUC of 0.749 and 0.624 for in-hospital and 1 year mortality, respectively.

Given that the hyperparameters were optimized from per-note topic features (that are themselves the output of an unstructured learning problem), it is most sensible that the topics information should be used in combination with the MTGP hyperparameters to describe patient state. We obtained improved predictive performance for both mortality outcomes when combining both MTGP hyperparameters with SAPS-I and the significant topics. This is likely because the hyperparameters provide complementary information to both SAPS-I and the significant topics. Both SAPS-I and the topic features capture a single aggregate measure of membership in certain latent dimensions related to outcome, while the MTGP hyperparameters capture movement over the course of a hospital stay within those dimensions. The best predictive performance occurred when all features were combined, e.g. SAPS-I + significant topics + MTGP hyperparameters.

3.4 Discussion

3.4.1 Using Aggregated Clinical Note Topics

Models that incorporated latent topic features were generally more predictive than those using only structured features, and a combination of the two feature types performed best. Notably, the combination provides a robustness that is able to perform well initially, leveraging primarily the structured information, and then continues to improve over the first 24 hours by incorporating the latent topic features. This resilience is particularly important since we observed that the first 24 hours of clinical notes appear to be the most meaningful toward predicting in-hospital mortality, while the predictive value of the baseline begins to steadily decrease. Note that similar features were also useful in predicting psychiatric readmission. [86]

Our observation of the importance of early data agrees with other reported results. Recall that, using topics derived from the first 24 hours of notes only, Lehman et al. obtained an average AUC for in-hospital mortality prediction of 0.78 (± 0.01), and this was increased to 0.82 (± 0.003) with the SAPS-I variable. Further, Hug et al. obtained an AUC of 0.809 for in-hospital mortality prediction based on information during the first 24 hours of ICU. As such, we examined our results for in-hospital mortality when using topics derived from the first 24 hours of notes only (prediction time of 36 hours in Figure 3-3), and obtained corresponding AUCs of 0.77 for the *Time-varying Topic Model*, and 0.841 for the *Combined Time-varying Model*. Compared to Lehman et al.'s result, this implies that (with enough data) neither the extra hierarchical learning nor the knowledge-based cleansing of medical terms before modeling improve prediction results (i.e., an AUC of 0.78 vs. 0.77). Compared to Hug et al.'s results, this implies that the addition of clinical text provides reasonable performance boosts to the power of gold-standard structured information like SAPS-II score (i.e. an AUC of 0.809 vs. 0.841).

Further, when predicting in-hospital mortality, we observed that the *Admission Baseline Model*'s predictive power (i.e., information acquired on admission) becomes much less valuable to predicting mortality as patients stay longer. This is likely because those who are not discharged within the first day of hospital admission are significantly sicker than those who are. Note that the average ICU stay time in the MIMIC II database is 3 days, and Figure 3-3 shows that after this time there was no additional predictive power gained by adding the structured admission information to the latent topic features (i.e., the *Time-varying Topic Model* and the *Combined Time-varying Model* converge).

This convergence draws attention to another interesting observation. Namely, both of the *Time-varying* models trended up in their ability to predict in-hospital mortality until 120 hours, and then trended down until the end of prediction. While initially counterintuitive, this is likely due to the loss of a significant number of patients (from both death and discharge) in the available patient cohort. For example, the test set population goes from 4,030 patients (3,626 control/404 positive for in-hospital mortality) to 3570 patients at this point (3,210 control/360 positive for in-hospital mortality).

Additionally, the predictive power of each topic changed depending on the target outcome. This appeals to intuition as, in a modern ICU, conditions that lead to in-hospital mortality are very different from those that would allow for a live discharge leading to a 30 day or 1 year mortality. As such, information about which topics tend to bias a patient towards any set of outcomes in useful for clinicians, when compared to the typical "black-box" approach to feature selection.

Finally, much work focuses on retrospective prediction of mortality outcomes. We also performed these predictions to compare the relative predictive power of different feature types and were able to achieve retrospective AUCs of 0.9, 0.94 and 0.96 for in-hospital mortality prediction using the *Retrospective Derived Feature Model*, *Retrospective Topic Model*, and combined *Retrospective Topic + Dervied Features Model.* However, we re-emphasize that predictions of mortality with retrospective feature sets are not helpful or relevant for clinical staff because statistical functions of signals or features (e.g., min/max) and other structured data (such as ICD-9 codes and EH comorbidities) are not known a priori.

3.4.2 Incorporating Features of Inter/Intra-signal Movement

The main limitation in using this approach to characterize timeseries is computational cost. We conducted an exhaustive grid search over the constrained hyperparameter space. We also used the NLML for the selection of the optimal hyperparameters of the MTGP, which may be sensitive to parameter initialization (due to the non-convex nature of this optimization function). Computational costs may be addressed using a recently proposed Bayesian optimization for automatically tuning the MTGP hyperparameters [98] in large datasets. In a "real-time" setting, the computational cost for m tasks is $O(m^3n^3)$. An overview of sparse GP methods is presented in [80], which aims to find a smaller set of pseudo-inputs n' to reduce computational complexity. In [1,7], some of these techniques were used to investigate sparse MTGPs, which reduce the complexity to $O(mnn'^2)$. Further improvement is possible by 1) exploiting the Kronecker product [96], 2) limiting the training data to the same time instances of each dimension of the data [24], or 3) by using recursive algorithms [77]. Applications that require close-to-real-time retraining (e.g., Experiment 2), would benefit from these techniques, while operating over longer time-scales would be less sensitive.

Further, in our approach the tasks are modeled with the same hyperparameters θ_t. Individual temporal covariance functions k_t for each task can be introduced using the idea of convolving two covariance functions, which has been described in [40] and further discussed in [66]. Our choice was motivated by the lower number of hyperparameters that have to be learned, and the concern that the introduction of convolved kernels may be inappropriate for real-world applications without a proper optimization process [40, 66].

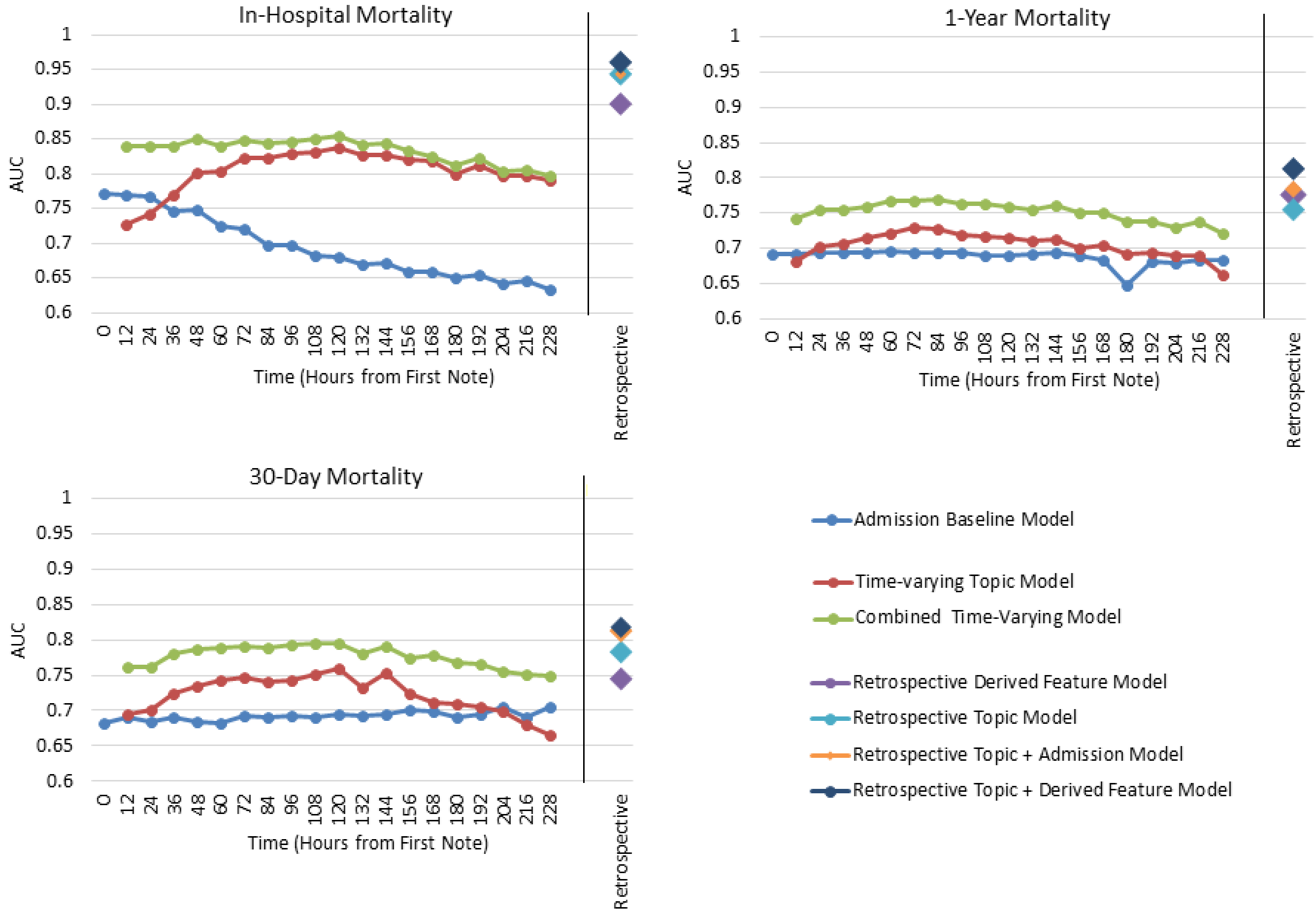


Figure 3-3: Linear SVM model performance measured via AUC on three outcomes: in-hospital mortality, 30 day post-discharge mortality, and 1 year post-discharge mortality. In each case, the features used are described in detail in Section 3.2.6. Our prediction task is different from the usual situation where data is accumulated over time. Since fewer patients have long ICU stays, in this case, we actually lose data points as time goes on, making the prediction task harder. For example, at time 0 there are 5,784 patients (5,157 controls/627 positives for in-hospital mortality) in the test set. By 72 hours, this had dropped to 5,084 patients (4,591 controls/493 positives for in-hospital mortality) and at 144 hours to 3,496 patients (3,141 controls/355 positives for in-hospital mortality). (Table A.1)

Table 3.3: Detailed model prediction results for three outcomes: in-hospital mortality, 30 day post-discharge mortality, and 1 year post-discharge mortality. This also appears in Figure 3-3.

Outcome Predicted	**Model Used**	**AUC**	**Sens.**	**Spec.**
In-Hospital Mortality	Admission Baseline Model	0.771	0.999	0.010
	Time-varying Topic Model 1	0.728	0.858	0.471
	. . .		. . .	
	Time-varying Topic Model 10	0.838	0.686	0.829
	. . .		. . .	
	Time-varying Topic Model 20	0.791	0.525	0.853
	Combined Time-varying Model 1	0.840	0.638	0.85
	. . .		. . .	
	Combined Time-varying Model 10	0.854	0.666	0.844
	. . .		. . .	
	Combined Time-varying Model 20	0.798	0.299	0.950
	Retrospective Derived Features Model	0.901	0.997	0.108
	Retrospective Topic Model	0.944	0.856	0.892
	Retrospective Topic + Admission Model	0.944	0.821	0.910
	Retrospective Topic + Derived Features Model	0.961	0.915	0.870
30 Day Mortality	Admission Baseline Model	0.683	0.995	0.075
	Time-varying Topic Model 1	0.695	0.150	0.944
	. . .		. . .	
	Time-varying Topic Model 10	0.759	0.817	0.551
	. . .		. . .	
	Time-varying Topic Model 20	0.665	0.602	0.579
	Combined Time-varying Model 1	0.761	0.348	0.885
	. . .		. . .	
	Combined Time-varying Model 10	0.796	0.641	0.770
	. . .		. . .	
	Combined Time-varying Model 20	0.75	0.011	0.991
	Retrospective Derived Features Model	0.745	0.941	0.220
	Retrospective Topic Model	0.783	0.342	0.909
	Retrospective Topic + Admission Model	0.813	0.872	0.633
	Retrospective Topic + Derived Features Model	0.818	0.096	0.985
1 Year Mortality	Admission Baseline Model	0.692	0.997	0.021
	Time-varying Topic Model 1	0.681	0.218	0.907
	. . .		. . .	
	Time-varying Topic Model 10	0.715	0.321	0.870
	. . .		. . .	
	Time-varying Topic Model 20	0.662	0.834	0.379
	Combined Time-varying Model 1	0.743	0.705	0.665
	. . .		. . .	
	Combined Time-varying Model 10	0.760	0.512	0.812
	. . .		. . .	
	Combined Time-varying Model 20	0.722	0.451	0.804
	Retrospective Derived Features Model	0.776	0.999	0.045

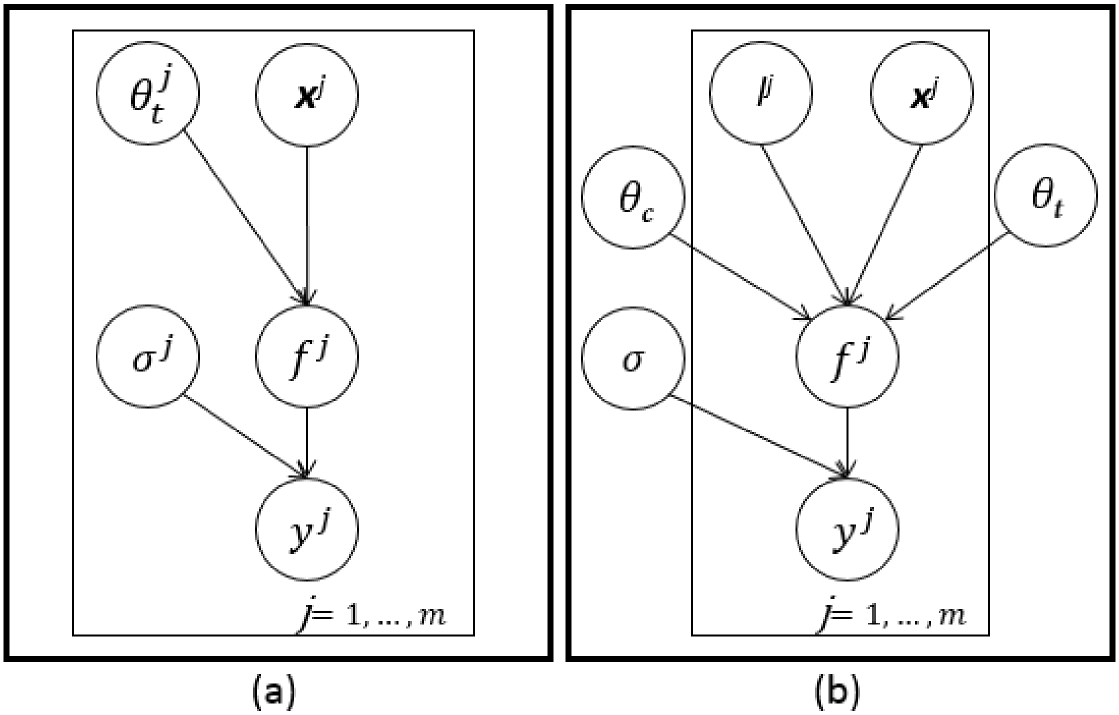


Figure 3-4: Graphical model for **(a)** m single-task Gaussian processes with m sets of: inputs X^i, temporal covariance hyperparameters θ_t^i, estimated functions f^i, noise terms σ^i, and outcomes y^i; and **(b)** a multi-task Gaussian process which relates m tasks through all prior variables, with the tasks' labels l and similarity matrix θ_c.

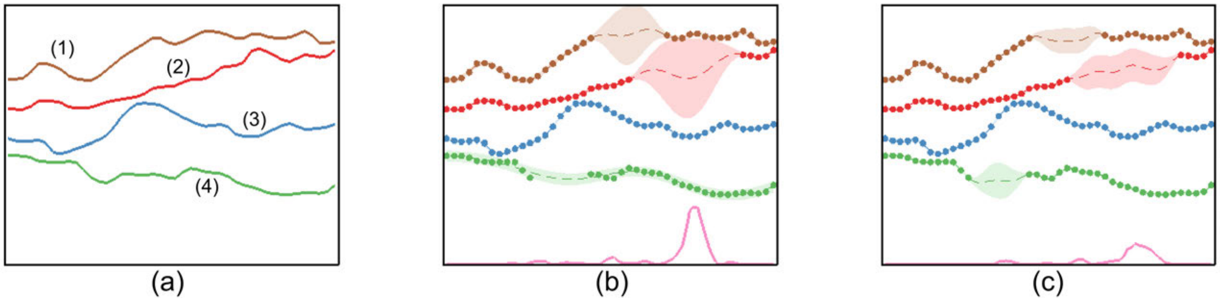


Figure 3-5: (a) A sample function with 4 tasks; (b) Single-task GP (STGP) and (c) multi-task GP (MTGP) predictions on all tasks. The dots represent observations, while dashed lines and colored areas represent the predictive mean and 95% confidence interval, respectively. The line on the bottom represents the mean absolute error (over the 4 tasks) between the predictions and the correspondent reference values. We observe that the overall error obtained in (c) is lower than that in (b), which suggests that the use of MTGP yielded better predictions by taking into account the correlation between the different tasks.

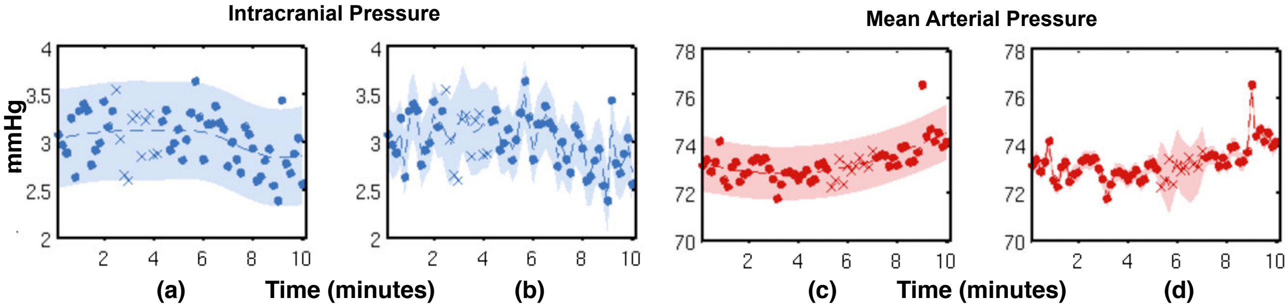


Figure 3-6: An example of a single-task GP (STGP) and multi-task GP (MTGP) applied to intracranial pressure (ICP) and mean arterial blood pressure (ABP) signals from a traumatic brain injury patient. (a) and (c) show the performance of STGP, whereas (b) and (d) show the improved performance of MTGP, which takes into account the correlation between ICP and ABP. Dots represent observations, crosses represent missing observations (test observations), the dotted line shows the function mean and the shaded area show the 95% confidence interval. We note that the timescale parameter "selected" by the MTGP, which takes into account the correlation between the tasks, is shorter than the one selected by the STGP, which yields to higher likelihood of the test observations (crosses).

Signal	Measure	STGP	MTGP
ICP	RMSE	0.91	0.69
	MSLL	0.6	0.45
ABP	RMSE	2.77	1.98
	MSLL	0.65	0.55
PRx-PRx*	RMSE	-	0.09

Table 3.4: Performance of single-task GP (STGP) and multi-task GP (MTGP). PRx-PRx* refers to the difference between the reference PRx (Pearson correlation coefficient of ICP and ABP for a given window) and PRx*, the estimated PRx index (posterior MTGP hyperparameter that measures the interaction between the two tasks).

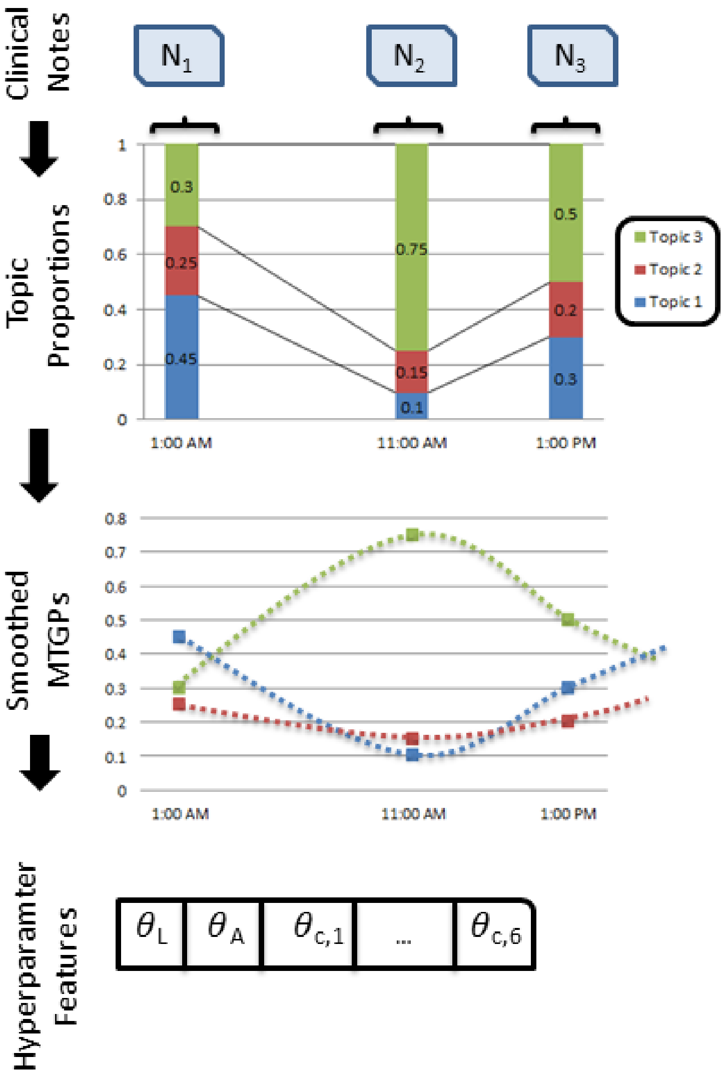


Figure 3-7: 1) We perform a pre-projection step where clinical notes are transformed into timeseries using Latent Dirichlet Allocation; 2) the new set of topic proportion timeseries are fitted using the MTGPs; 3) inferred hyperparameters $\theta_L, \theta_A, \theta_{c,1}, \ldots, \theta_{c,6}$ are derived, projecting into the new latent space; 4) latent features (hyperparameters) are used as features in combination with topic proportions and the SAPS acuity score to 5) forecast patient mortality.

	Top Five Words	**Possible Topic**
In-hospital Mortality	liver, renal, hepatic, ascites, dialysis	Renal Failure
	thick, secretions, vent, trach, resp	Respiratory infection
	remains, family, gtt, line, map	Systematic organ failure
	increased, temp, hr, pt, cc	Multiple physiological changes
	intubated, vent, ett, secretions, propofol	Respiratory failure
	name, family, neuro, care, noted	Discussion of end-of-life care
Survival	cabg, pain, ct, artery, coronary	Cardio-vascular surgery
	chest, pneumothorax, tube, reason, clip	
	pain, co, denies, oriented, neuro	Responsive patient

Table 3.5: Top five words in chosen topics (enriched for in-hospital mortality/survival).

Features	Hospital Mortality	1-Year Mortality
SAPS-I	0.702	0.500
Ave. Topics	0.759	0.653
SAPS-I + MTGP	0.775	0.624
Ave. Topics + MTGP	0.788	0.673
SAPS-I + Ave. Topics + MTGP	0.812	0.686

Table 3.6: Prediction results of hospital and 1-year mortality, AUC, for Gaussian process feature combinations.

Chapter 4

Voice Disorder Detection in Wearable Out-patient Devices

4.1 Background

An estimated 7% of the working-age population in the U.S. is affected by a voice disorder [69, 85]. Most cases of voice disorders result from vocal misuse (exerting excessive muscle force or physical effort while vocalizing). This is typically referred to as *vocal hyperfunction*. We define "vocal hyperfunction" to refer to patterns of vocal behavior that could be harmful. Vocal hyperfunction is not always present, and therefore may not exhibit in clinic. In some patients vocal hyperfunction causes a deterioration in voice quality and vocal fatigue but without any underlying tissue pathology; this is commonly referred to as *muscle tension dysphonia* (MTD). Unlike those with vocal fold pathology (e.g. nodules or polyps), MTD patients are notoriously difficult to characterize because there is no consensus on an objective biomarker. Previous studies have also demonstrated that commonly held "indicators" of MTD appear frequently in individuals who have no known voice disorder [3, 94].

Because MTD is behaviorally induced, treatment typically involves an attempt to modify vocal behavior through speech/voice therapy [42]. However, MTD can be manifested in a wide range of maladaptive vocal behaviors (e.g., various degrees of strain or breathiness) whose nature and severity can display significant situational variation (e.g., variation associated with changes in levels of stress [21]). Clinicians currently rely on patient self-reporting and self-monitoring to assess the prevalence and persistence of these behaviors during diagnosis and management. But these reports are highly subjective and are known to be inaccurate [9, 71, 82].

4.2 Overview

The work reported here is part of an ongoing project to gain insight into the complex relationships underlying vocal hyperfunction by analyzing data collected from an accelerometer (ACC) placed on the neck [65]. We use an accelerometer rather than an acoustic microphone to protect the privacy of subjects. Recent studies have demonstrated some success applying supervised learning to ACC data to distinguish between patients with and without existing vocal fold pathology [34]. In this work, our goal was to determine if specific patterns of glottal pulses were associated with MTD, a type of non-phonotraumatic hyperfunction. This is more challenging in three respects:

- Patients with muscle tension dysphonia (MTD) have a behavioral disorder whereby they misuse their vocal folds, but do not have an anatomical abnormality. Therefore their voices are sometimes abnormal and sometimes not.
- While it is possible to obtain subjective expert-generated labels for acoustic recordings, it is impossible to obtain labels at the level of individual utterances for hundreds of millions of utterances. Additionally, even if someone were willing to devote the time to labeling a substantial number of utterances, the mapping between the ACC signal and voice misuse is not currently known. Consequently, there is no opportunity to use supervised learning to classify utterances.
- Rather than attempting to classify individual subjects, we attempt to uncover the key differences between many kinds of intermittently occurring hyperfunctional and normal voice use—without prior knowledge of what characterizes such behaviors.

We attack the problem of quantifying vocal hyperfunction by clustering glottal pulses using symbolic mismatch [100]—a technique previously used to study ECG signals. We segmented over 110 million glottal pulses from the ACC signals for subjects, and then clustered them into symbols. We then used symbolic mismatch to compare the frequencies and shapes of those symbols between subjects, leading to a distance measure between each pair of subjects. Finally, based on this distance measure, we clustered subject-days.

To evaluate our approach, we used 253 subject-days of data obtained from 11 patients and 11 matched controls (*Control*). Data from patients was gathered both before they underwent voice therapy (*PreTx*) and after voice therapy (*PostTx*). Though we know that each individual exhibits different vocal behaviors within a day, we hypothesized that subject-class-specific differences in the distribution of the behaviors would be reflected in the distribution of subject days in each cluster. To check this we calculated a total concentration measure based on the density of each class of subject in each cluster.

Devices that use a neck-placed miniature accelerometer (ACC) as a phonation sensor have shown potential for accurate, unobtrusive, and privacy-preserving long-term monitoring of vocal function [65] (Figure 4-1). The individual periods (pulses) in the ACC signal have a general shape that reflects the vibratory pattern of the vocal folds during phonation, and vary with changes in vocal function/quality. Recently, researchers have examined vocal hyperfunction using summary features obtained from ambulatory monitoring [34, 84], but these assessments were based on aggregates, and were not designed to detect periods of hyperfunction. Glottal pulses obtained from the ACC signal have a general shape that describes the acceleration of the vocal folds as they vibrate to create airflow for voice production. Because ACC signals have only recently become available, variations in the segmented pulses are not currently well-characterized.

4.3 Methods

To generate symbols for every subject-day tuple, we segmented each daily ACC signal into non-overlapping frames to create a set of variable-length, peak-to-peak glottal pulse segments. We then computed the pulse-to-pulse distance using a lower bounds to dynamic time

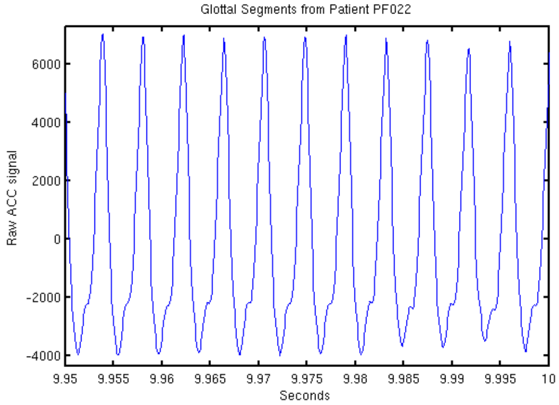


Figure 4-1: A sustained vowel "a", containing 10 peak-to-peak glottal pulses in 0.05 seconds.

warping (DTW) distance, and created clusters iteratively as described below.

4.3.1 Glottal Pulse Symbolization

Segmentation We begin with the continuous univariate timeseries of a single subject's ACC recording on a given day (a "subject-day"). This signal $\mathbf{x} \in R^T$ is a collection of T samples, i.e. $\mathbf{x} = \{x_1, ..., x_t, ..., x_T\}$, in which measurements are regularly-sampled. We split the ACC signals into individual glottal pulses by detecting characteristic peaks. Peak detection involved 1) using an off-the-shelf peak detection algorithm [63] to make a first guess at peak locations based on amplitude, and 2) using an estimate of the subject's underlying vocal pitch to correct missing and spurious peaks. After segmentation, we have a vector of M daily glottal pulse segments, $\mathbf{x}_{\text{seg}} = \{(x_{t_1}, ..., x_{t_2}), ..., (x_{t_{2M-1}}, ..., x_{t_{2M}})\}$, where $t_1, ..., t_{2M}$ are increasing but not necessarily contiguous, so that $0 \leq t_1 \leq ... \leq t_{2M} \leq T$. Notationally, we re-label this as $\mathbf{x}_{\text{seg}} = \{z_1, ..., z_M\}$ where $z_1 = (x_{t_1}, ..., x_{t_2})$, $z_M = (x_{t_{2M-1}}, ..., x_{t_{2M}})$.

The amplitude of each glottal pulse was scaled to units of sound pressure decibels (dbSPL) based on an estimated linear fitting between ACC signal units and average dbSPL for the subject on that day to determine periods of voicing. The length of each individual segmented pulse varied; to compare all pulses, we length-normalized pulses by evenly up-sampling all segments to the longest segment length.

Pulse-Pulse Distance Computation Silent segments were grouped by their length into bins of 1 second, 1 minute, 10 minutes, and an hour or more.[1] To account for the large variation in subjects' patterns of voice use across days (e.g., teachers typically spoke less on weekends), we chose to examine each day separately. For each subject-day, we start with the constructed vector $\mathbf{x}_{\text{seg}} = \{z_1, ..., z_M\}$ and compute the distance between all pulses z_i and z_j using the Keogh Bounds (LB_Keogh) [107] as a surrogate for DTW. LB_Keogh is a tight lower bound to DTW between a candidate signal C and query signal Q, and is considerably more computationally efficient than DTW.

Symbolization for Symbolic Feature Creation We next used hierarchical clustering with Ward's linkage, which minimizes the total within-cluster variance, to cluster a randomly selected initial subsample of 3,000 pulses per subject-day. We used a distance cutoff of 30% of the maximum distance to determine k, the number of clusters. Having chosen k, we then used iterative k-means to cluster all of the pulses $z_1, ..., z_M$. Each of the k clusters can be considered as representing a class of glottal pulses whose members have a similar shape. We label each of these classes with its centroid, and create a vector of length k of symbolic features $\mathbf{v}$ for each subject-day, where $\mathbf{v} = \{(s_1, f_1)..., (s_k, f_k)\}$, s_i is the i^{th} class centroid, and $f_i = \frac{|s_i|}{\sum_j |s_j|}$. In creating $\mathbf{v}$, we have now abstracted from a stream of millions of glottal pulses into a finite alphabet of symbols with matching frequencies of occurrence.

Symbolic Mismatch Distance Measure Once symbolic features $\mathbf{v}$ were created for each subject-day, we defined the overall distance measure between each pair of $\mathbf{v}$'s as the symbolic mismatch distance D_{mismatch}[i,j]. For subject-days $\mathbf{v}_i$ and $\mathbf{v}_j$, D_{mismatch}[i,j] is the aggregate sum of the weighted distance between class centroids.

[1] A lot of any subject's day is spent in silence; the amount varied from 86%-95%. The mean number of voiced pulses per patient was 3,427,367.

Algorithm 1 Symbolic mismatch calculation between subject/day tuple pairs.

```
Input: Transformed data from subject/day tuples $v_i$ and $v_j$
Output: Weighted distance between $v_i$ and $v_j$
initialize W ← 0
for each $s_a \in v_i$ do
    for each $s_b \in v_j$ do
        W ← W + $f_a$ * $f_b$ * LB_Keogh($s_a$, $s_b$)
    end for
end for
$D_{\text{mismatch}}$[i, j] ← W
```

4.3.2 Subject-Day Clustering and Evaluation

We evaluate a clustering of Q subject-days $\mathbf{v}_1, ..., \mathbf{v}_Q$ across n clusters in two ways: *class concentration* and *subject concentration.* For an individual cluster $\mathbf{c}$ with some number of total (subject-day, class label) pairs, i.e., suppose there are o pairs of them $\mathbf{c} = \{(\mathbf{v}_1, \mathbf{l}_1), ..., (\mathbf{v}_o, \mathbf{l}_o)\}$, class concentration is the cluster's ratio of the dominant label to the total number of in-cluster subject-days. Subject concentration is calculated similarly, but we count $\mathbf{v}$ from the same subject only once. For example, suppose we have a cluster with items $c_1 = \{(v_{1-1}, 0), (v_{2-1}, 1), (v_{2-3}, 1), (v_{3-1}, 1), (v_{3-5}, 1)\}$[2], the class concentration would be $conc_{class} = \frac{4}{1+4}$ and the subject concentration would be $conc_{subj} = \frac{2}{2+1}$.

Total Concentration We define the total concentration for both metrics across clusters as the weighted sum of all individual cluster concentrations. Specifically, for n clusters $c_1 \dots c_n$ with concentrations $h_1 \dots h_n$, total concentration is defined as $total\_conc = \sum_{i=1}^{n} h_i * |c_i|$. Note that when there are two classes, the total concentration can range from $[0.5, 1]$, since the least concentrated cluster possible is 0.5. To check statistical significance, we tested the null hypothesis that the groupings obtained with D_{mismatch} were different from a total concentration measure using random distances. We first define a random distance metric (RRDM) by sampling random values uniformly as $RRDM[i, j] = \mathcal{U}([0, max\{D_{\text{mismatch}}\}])$, where $max\{D_{mismatch}\}$ is the maximum distance seen from the actual symbolic mismatch.

[2] Corresponding to subject 1-day 1 with label 0, subject 2-day 2 and subject 2-day 3 labeled 1, etc.

We sampled distances for each subject/tuple pair $\mathbf{v}_i$ and $\mathbf{v}_j$ 5,000 times, and cluster those (random) values. We clustered the RRDM values to obtain a distribution of total class concentration measures, fit an empirical CDF (ECDF) to these values, and computed the probability (p) of a total class concentration value greater than or equal to ours by chance $(1 - ECDF(conc_{class}(D_{\text{mismatch}})))$.

4.4 Experiments

4.4.1 Data

We considered 11 MTD patients with matched controls — a total of 22 subjects. Diagnoses were based on evaluation by a laryngologist and speech-language pathologist. All patients were treated with behavioral voice therapy, and each patient was recorded for a minimum of six days both before and after undergoing treatment. This created a set of three categories in our data:

- 11 pre-treatment MTD patients **(PreTx)**,
- the same 11 patients after behavioral voice therapy **(PostTx)**, and
- 11 control subjects matched for age, gender, and occupation **(Control)**.

We used a neck-placed miniature accelerometer as a voice sensor and a smart phone as the data acquisition platform [65]. The raw accelerometer signal was collected at 11,025 Hz, 16-bit quantization, and 80-dB dynamic range in order to obtain neck skin vibrations at frequencies up to 4,000 Hz. Our dataset contains 253 subject-days, corresponding to over 110 million segmented pulses (details in Appendix A). Working with a continuous ACC signal for each subject over the course of 7+ days yielded approximately 15 GB of data per subject.

4.4.2 Clinical Significance

We investigated the utility of our method in addressing three clinical questions:

1. **Can our features be used to diagnose MTD (PreTx vs. Control subject/days)?** To address the first question, we performed an inter-subject comparison on PreTx vs. Control subjects, where we clustered all pre-therapy subject-days and all control subject-days. We did not expect a clean separation of all PreTx days from Control days to occur, because many MTD patients have "good" days where their voice use is like that of a vocally normal individual. Instead, our objective was to determine if a clustering of D_{mismatch} could achieve a high concentration in the PreTx vs. Control comparison ($conc_{class}(PreTx/Con)$) that was significantly different from those that could be obtained by chance.

2. **Can we detect a treatment effect (paired PreTx vs. PostTx subject/days)?** To address the second question, we perform an intra-patient comparison on PreTx vs. PostTx subjects where we performed clusterings on a patient-patient basis, (i.e., we clustered all days, both pre and post treatment, on a patient-by-patient basis).

3. **If our features can be used to detect treatment effect, is the effect to move patients towards "normal" (PostTx vs. Control subject/days)?** To address this question, we performed an inter-subject clustering on the PostTx vs. Control subjects, clustering all post-therapy subject-days and all control subject-days. Our objective was to determine if this clustering would produce concentrations ($conc_{class}(PostTx/Con)$) which were not significantly different from those that could be obtained by chance. This would indicate that patients are difficult to distinguish from controls after they receive voice therapy.

4.4.3 Baseline Methods

Our symbolic features (SF) were compared over subject-days using symbolic mismatch to generate a paired distance matrix, and the mismatch distance was clustered using hierarchical clustering and Ward's linkage. We compared clusterings generated from our method to clusterings from features generated by a recently proposed system for identifying phonotraumatic hyperfunctional patients with pathology (nodules or polyps) versus their matched controls [34].

As in [34], we windowed the regularly sampled $\mathbf{x} = \{x_1, ..., x_t, ..., x_T\}$ ACC signal into five-minute windows, computed the phonation frequency (f0) and acoustic sound pressure level (SPL) of non-overlapping 50 millisecond frames within each window (i.e., 6000 frames per window), and extracted statistical features of these acoustically inspired measures (e.g., the mean, skew, 5^{th} percentile value, etc.). Each subject-day is a feature matrix, where the number of features varied based on the amount of phonation in each subject-day. We also removed the most correlated features, yielding a total of 22 features. Once generated for each subject-day, these generate a Vector of Acoustic Features (VAF) that has multiple features summarizing a given subject-day tuple. We clustered VAF vectors from all subject-day tuples using k-means clustering with a squared Euclidean distance function.

While the VAF previously detected constantly-present pathology in phonotraumatic patients, we theorized they would create many incorrectly labeled windows for clustering in the periodically hyperfunctional MTD population. To address this, we took the feature-wise mean over all five-minute windows for a single subject-day, to obtain Mean Acoustic Features (MAF). These vectors were clustered with hierarchical clustering and Ward's linkage.

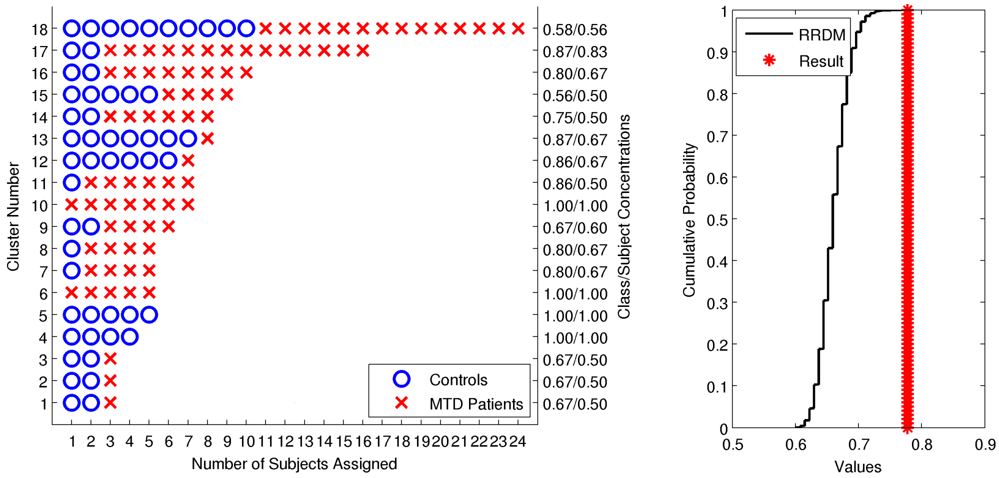
We measured the total concentration in all clusterings as described in 4.3.2. For inter-subject comparisons, we investigated the sensitivity of our method and the baselines by varying the number of clusters in the final grouping (n) from 2 to 40; for the intra-subject comparisons we varied n from 2 to 10.

4.5 Results

4.5.1 Control vs. PreTx Subjects—Potential for ambulatory screening tool

After performing clustering on all subject-day pairs from Control and PreTx subjects into 18 clusters, we obtained a total class concentration measure of 0.70. As shown in Figure

4-2, using the RRDM clustering comparison, the difference between the PreTx and Control groups was statistically significant at $p < 0.001$. There were a total of 135 subject/days in the comparison, and no cluster had data from only a single subject (total subject concentration measure of 0.65). Given the intermittent nature of voice misuse, it is reasonable that some days from PreTx patients cluster with Controls.



(a) Clustering of PreTx vs. Control subject-days

(b) Result vs. RRDM ECDF

Figure 4-2: We show a) the results of symbolic mismatch clustering of the control subject-days versus the PreTx subject-days (18 clusters, class concentration = 0.70) and b) the empirical CDF of the 5,000 RRDM clusterings versus our experimental results ($p = 0.001$). As shown in b), Controls and PreTx patients were significantly different.

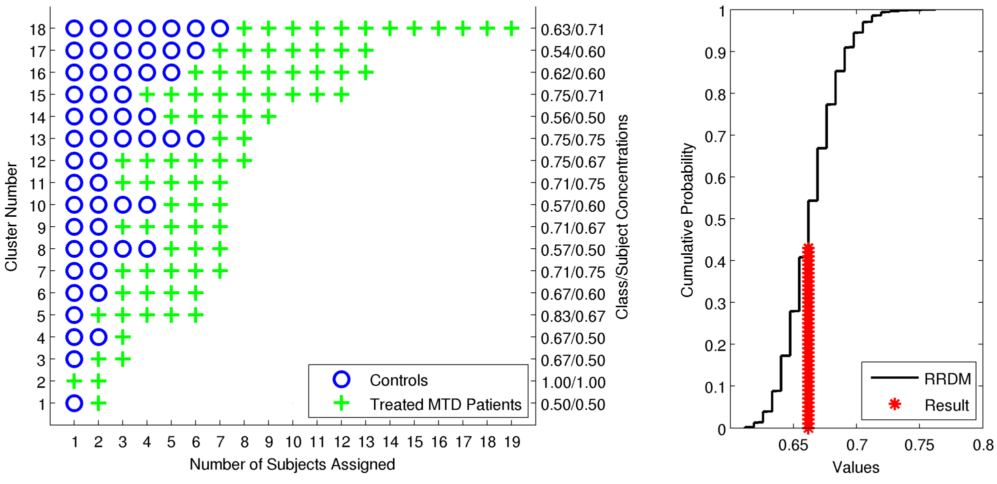
4.5.2 PreTx vs. PostTx Subjects—Vocal therapy effect in pairs

We investigated if voice therapy had an effect that could be detected in our framework by using an intra-subject comparison on a patient-patient basis, so that all days from a patient pre-treatment were compared all days from the same patient post-treatment. As shown in Table 4.1, the results vary for each patient, with some demonstrating more post-therapy differences than others. One possible explanation for a smaller intra-subject concentration is that improved vocal behavior for a particular subject was observable during a smaller time scale than we examined (e.g., better behavior during their evenings).

Table 4.1: Total concentration of per-patient PreTx vs. PostTx with three clusters. Concentrations that passed the empirical RRDM significance of $p < 0.01$ are highlighted with **, and those with $p < 0.05$ are marked with *.

F023	F027	F040	F048	F052	F064	F069	F071	F100	M035	M074
0.73	0.65	0.81*	1.0**	0.63	0.69	0.67	1.0**	0.86*	0.57	0.79

4.5.3 Control vs. PostTx Subjects—Therapy moves subjects toward "normal"



(a) Clustering of PostTx vs. Control subject-days

(b) Result vs. RRDM ECDF

Figure 4-3: We show a) the results of symbolic mismatch clustering of the control subject-days versus the PostTx subject-days and b) the empirical CDF of 5,000 random distance clusterings versus our experimental results. PostTx subject-days were not significantly different from the control group, suggesting that voice therapy does indeed move patients toward vocal normalcy.

As shown in Figure 4-3, after clustering the PostTx patients and Control subjects, we obtained a total class concentration of 0.63, and a subject concentration of 0.60. There was no statistically significant difference between these clustering and clusterings of the RRDM distances ($p = 0.56$). In this clustering of the 139 total days, PostTx patients only enrich a few clusters, and many clusters are evenly class-balanced. This suggests that our method is picking up changes caused by voice therapy, and that these changes are in the right direction.

4.5.4 Sensitivity Analysis of Clustering Across Baselines and Clusters

After successfully demonstrating differences in PreTx vs. Control subjects-days, and showing that PostTx subject-days are like those of the Controls, we examined the ability of our symbolic features (SF) to perform under varying numbers of clusters as compared to other methods (VAF and MAF).

We first computed the concentration values for which RRDM passes the p < 0.01 significance level; our SF features should ideally keep the total concentration of the PreTx/Control clustering over p < 0.01, and the PostTx/Control clustering under p < 0.01 to demonstrate that there are consistent differences from the Control subject-days in the PreTx group that are not present in the PostTx group after therapy. As shown in Figure 4-4, the inter-subject class concentration increases as the number of clusters grows. The Vector Acoustic Features (VAF) perform worst, followed by the Mean Acoustic Features (MAF). The MAF PreTx-Control nears statistical significance. With our method (SF) PreTx-Control clusterings are significant at the 0.05 level on all but the very first clusters. We also have the PreTx and PostTx group separate when more than 5 clusters are used, and the separation passes the RRDM p < 0.01 significance level. Specific clustering results for $n = 18$ ($d = 0.116$) are presented in Sections 4.5.1 and 4.5.3.[3]

4.6 Discussion

In this work we used unsupervised machine learning to analyze a novel clinical data set containing long-term time-series data. Prior work with ACC data has focused on targeted feature extraction for supervised classification of subjects [34]. However, supervised learning is a poor method for detecting differences in the vocal behavior of MTD patients, because people with MTD do not always speak in a disordered way, and there is no standard for

[3] The distance between the the PreTx and PostTx concentrations was maximized in our method when 24 clusters were used (total class concentration difference = 0.124). However, $n = 18$ minimized the number of clusters over the maximum concentration difference $d = conc_{class}(PreTx/Control) - conc_{class}(PostTx/Control)$, such that d was not significantly lower than the absolute $\max_{n}(d)$.

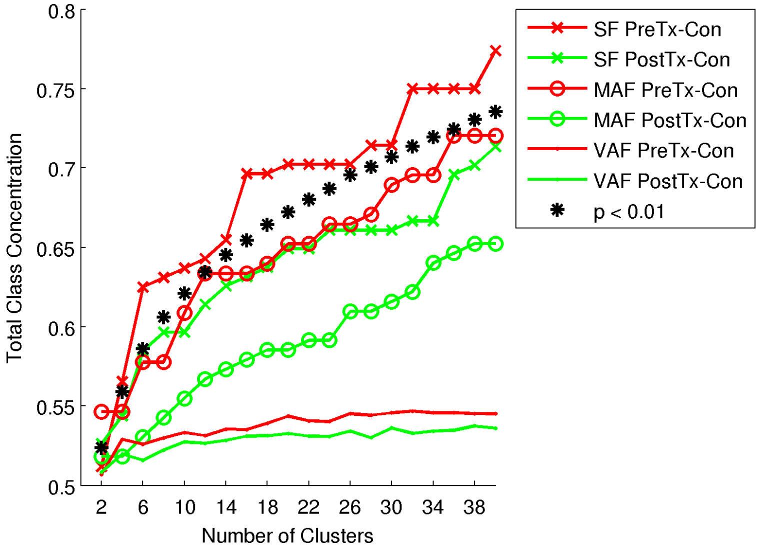


Figure 4-4: The sensitivity of inter-subject clustering results for VAF, MAF and SF methods. The PreTx group is consistently more concentrated than the PostTx group for all methods, but only our method demonstrates the SF PreTx/Control clustering passing statistical significance.

labeling individual glottal pulses as disordered.

Our method differs from other recent work in three key ways: 1) We segment individual glottal pulses from the ACC signal rather than taking the traditional fixed-width frames; 2) We directly judge the relevance of a particular segmented item in our set by its morphology rather than using transforms derived from expert knowledge; and 3) We summarize a subject-day using a weighted sum over paired sets of morphological symbols and frequencies rather than a large set of features, or simple aggregates. From a clinical perspective, our results demonstrate that an ACC signal can be used to detect a difference in the vocal behavior of patients and controls. We also showed that vocal therapy has a measurable impact on patient behaviors.

Time-series symbolization [58] and symbolic representation for time series based on sequence shape [74] have previously been used to find time series motifs. Symbolization of segmented ECG data was used for supervised risk stratification [99] and assessing the clinical utility of expert-annotated heartbeats [57]. Unlike this prior work, we do not use symbolized distances as part of a supervised learning regime. Instead, we use these distances to represent using a set of density-based prototypes.

More complex generative models have recently been developed for physiological problems, e.g., using a multi-level latent model to learn individual and population level traits from clinical temporal data [91].Symbolization is particularly attractive for developing clinical markers, since symbols are fast to extract and compare [100], and variations in glottal pulse shape based on voice quality may be detectable with symbolization [51].

Our study uses a week of data because individuals tend to have a regular schedule that changes over the course of a week. It is possible that less data is needed to understand voice misuse, however we believe generally that the minimum amount of time needed to cluster would vary strongly depending on the subject's degree of hyperfunction, and how regularly they misuse their voice. For example, a professor may only misuse her voice on Tuesdays and Thursdays because of the additional strain that a 2 hour lecture adds. Many standard clinical measures used to identify vocal hyperfunction do not fit the MTD patient's behavioral condition. This is likely why the VAF and MAF measures do not work well.

Our work is the first large scale study of vocal misuse based on long-term ambulatory data with over 100 million segments corresponding to glottal pulses from 253 subject-days of data. The long-term goal of this multi-disciplinary project is to build a non-invasive ambulatory system that could be used to 1) diagnose voice disorders, 2) assess the impact of voice therapy, and 3) help facilitate the adoption of more normal vocal behaviors by providing biofeedback.

Chapter 5

Conclusion

In this work, we focused on problems that require recognition of, and response to, patients in a physiologically compromising states. We specifically targeted the evaluation of representations of multi-modal clinical data that are useful for predicting important clinical tasks. In the future, we hope that such methods will be used to provide clinical staff with the support to improve decision making for patient care.

We covered work that spans coded records from administrative staff, vital signs recorded by monitors, lab results from ordered tests, notes taken by clinical staff, and accelerometer signals from wearable monitors. In tackling these problems, our focus was on learning abstractions that generalized across applications despite missing and noisy data. In general, our experimental process targeted learning techniques that transform diverse data modalities into a consistent intermediate that improves prediction in clinical investigation.

There are major practical and technical barriers to understanding human health. We believe that the creation of machine learning methods to distill large amounts of heterogeneous health data into evidence-based clinical support will advance scientific understanding, and we hope this will lead to improved human health.

There are many exciting opportunities for work in this vein to continue.

Early Prediction of In-Patient ICU Interventions Our work in this space [35, 36] focused on clinically actionable prediction tasks, with an emphasis on representations that improve task performance on multiple potential targets. We also wanted the latent states to qualitatively make sense for the task.

There are several limitations to this body of work. First, it was limited to ICU patients and there are interesting questions about its applicability to broader non-ICU settings, for example the emergency room or during outpatient treatments. There is also much uncertainty about the reason for specific clinical behaviors (e.g., why an early prediction of weaning is confirmed by notes, but did not occur until significantly later). This is difficult to quantify because we do not have ground truth on the decision making process for any ICU intervention.

Future directions for this work should include new way to model missingness. We chose physiological words, but other interesting approaches also exist. There are also many ways to have a learned representation capture higher-level structure and dependencies between multi-modal time series data and multiple time-varying targets. It is also an open question as to how one should best balance learning the distribution of the data with trying to push a discriminative signal into the learning process (combining supervised and unsupervised methods).

Representations for Post-Discharge Outcome Prediction This work [30,32] focused on learning representations of clinical notes that are predictive of in-hospital mortality and post-discharge mortality. Other work has validated the value of these representations on predicting psychiatric readmission [86]. Topic and kernel representations provide interesting spaces for intuitive comparisons of patients, and the representations improve task performance. There is also the promise of possible actionability for each of these tasks, for example by allocating home visits to patients who are likely to have a psychiatric readmission, or doing a hospice discharge for patients with a high risk of 30-day mortality.

In general, this work is limited by its use of mortality as a proxy for acuity in the ICU. In general, modern ICU patients do not die in-hospital unless clinicians turn off support. This may lead to "decisions" about patient care that may actually be made far before it seems that they have been, and our learning may actually fall temporally behind clinical decision processes. There are also general questions that we did not address about the scalability of kernel-based methods for larger datasets and the generalization of model-based methods to multi-center datasets. Finally, this work considered model prediction times that were fixed or batched, which may limit their usefulness in a more dynamic environment.

New work in this space should consider more diverse measures of severity of illness other than mortality. For example, one such measure might be the amount of deviation a patient has from past reported values, which can be considered their own "norm". There are also many choices for more robust note representations that also take note authors or types into account. and many kernels that may be better suited to various forms of intra/inter signal modeling. Finally, there is much value to combining data across modalities and time in this work as well.

Voice Disorder Detection in Wearable Out-patient Devices We focused in early work on predicting pathologies with acoustic-like features [34] for subjects with phonotraumatic voice disorders, and then on detecting harmful patterns with glottal-pulse based features [33] in subjects with non-phonotraumatic voice disorders. In our more recent work, we were most excited by the possibility of understanding what moves subjects towards âĂIJnormalâĂİ after voice therapy.

This work was limited first by the small size of our patient cohort, which is currently at 50 patients and their matched controls (100 total). While each subject generated a large amount of data, more work is needed to understand the generalizability of our findings. Further, the data from non-phonotraumatic subjects has no behavioral ground truth available—meaning that we do not truly know if a particular behavior most commonly found in a subject (but not in controls) is a "damaging" one. There is an additional chicken/egg problem in phonotraumatic subjects, where we only have subjects after they acquire pathology, and so we cannot know if features most predictive of pathology were in fact caused by it or were the cause of the pathology. Finally, post-hoc interpretation of the ACC signal is challenging, as it is not currently a widely-used signal.

Opportunities in this space for more progress should focus on other types of windowing for learning, particularly utterance-based analysis. For real understanding, there should also be testing of detected features in biofeedback studies to understand impact. There is also interesting new work in the non-clinical setting that should target the use of non-invasive wearable data to detect harmful behaviors in general. Important questions in this space are whether we can find meaningful behavioral needles in large haystacks of data.

Bibliography

[1] Mauricio A. Alvarez and Neil D. Lawrence. Computationally efficient convolved multiple output gaussian processes. *The Journal of Machine Learning Research*, 12:1459âĂŞ1500, 2011. 00031.

[2] C.W. et al. Arnold. Clinical case-based retrieval using latent topic analysis. In *AMIA Annual Symposium Proceedings*, volume 2010, page 26. AMIA, 2010.

[3] Alison Behrman, Linda D Dahl, Allan L Abramson, and Harm K Schutte. Anterior-posterior and medial compression of the supraglottis: signs of nonorganic dysphoniaor normal postures? *Journal of Voice*, 17(3):403–410, 2003.

[4] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[5] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *JMLR*, 3(5):993–1022, 2003.

[6] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

[7] Edwin V Bonilla, Kian Ming Adam Chai, and Christopher KI Williams. Multi-task gaussian process prediction. In *Advances in Neural Information Processing Systems*, pages 153–160, 2007.

[8] Michael J Breslow and Omar Badawi. Severity scoring in the critically ill: Part 1âĂŤinterpretation and accuracy of outcome prediction scoring systems. *CHEST Journal*, 141(1):245–252, 2012.

[9] R Buekers, E Bierens, H Kingma, and EHMA Marres. Vocal load as measured by the voice accumulator. *Folia phoniatrica et logopaedica*, 47(5):252–261, 1995.

[10] Karla L. Caballero Barajas and Ram Akella. Dynamically modeling patient's health state from electronic medical records: A time series approach. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 69–78, New York, NY, USA, 2015. ACM.

[11] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1721–1730. ACM, 2015.

[12] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM TIST*, 2:27:1–27:27, 2011.

[13] Dustin Charles, Meghan Gabriel, and Michael F Furukawa. Adoption of electronic health record systems among us non-federal acute care hospitals: 2008-2012. *ONC data brief*, 9:1–9, 2013.

[14] Zhengping Che, David Kale, Wenzhe Li, Mohammad Taha Bahadori, and Yan Liu. Deep computational phenotyping. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 507–516, New York, NY, USA, 2015. ACM.

[15] Zhengping Che, David Kale, Wenzhe Li, Mohammad Taha Bahadori, and Yan Liu. Deep computational phenotyping. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 507–516. ACM, 2015.

[16] D.A. Clifton, L. Clifton, S. Hugueny, D. Wong, and L. Tarassenko. An extreme function theory for novelty detection. *Selected Topics in Signal Processing, IEEE Journal of*, 7(1):28–37, Feb 2013.

[17] L. Clifton, D.A. Clifton, M.A.F. Pimentel, P.J. Watkinson, and L. Tarassenko. Gaussian processes for personalized e-health monitoring with wearable sensors. *Biomedical Engineering, IEEE Transactions on*, 60(1):193–197, Jan 2013.

[18] Mitchell J Cohen, Adam D Grossman, Diane Morabito, M Margaret Knudson, Atul J Butte, and Geoffrey T Manley. Identification of complex metabolic states in critically injured patients using bioinformatic cluster analysis. *Critical Care*, 14:R10, 2010.

[19] M Czosnyka, P Smielewski, P Kirkpatrick, RJ Laing, D Menon, and JD Pickard. Continuous assessment of the cerebral vasomotor reactivity in head injury. *Neurosurgery*, 41(1):11–17, 1997.

[20] Daniel De Backer, Patrick Biston, Jacques Devriendt, Christian Madl, Didier Chochrad, Cesar Aldecoa, Alexandre Brasseur, Pierre Defrance, Philippe Gottignies, and Jean-Louis Vincent. Comparison of dopamine and norepinephrine in the treatment of shock. *New England Journal of Medicine*, 362(9):779–789, 2010.

[21] L Demmink-Geertman and PH Dejonckere. Neurovegetative symptoms and complaints before and after voice therapy for nonorganic habitual dysphonia. *Journal of Voice*, 22(3):315–325, 2008.

[22] R Durichen, M Pimentel, Lei Clifton, Achim Schweikard, and D Clifton. Multi-task gaussian processes for multivariate physiological time-series analysis. 2014.

[23] Frederick DâĂŹAragon, Emilie P Belley-Cote, Maureen O Meade, François Lauzier, Neill KJ Adhikari, Matthias Briel, Manoj Lalu, Salmaan Kanji, Pierre Asfar, Alexis F Turgeon, et al. Blood pressure targets for vasopressor therapy: A systematic review. *Shock*, 43(6):530–539, 2015.

[24] Theodoros Evgeniou, Charles A. Micchelli, and Massimiliano Pontil. Learning multiple tasks with kernel methods. In *Journal of Machine Learning Research*, page 615âĂŞ637, 2005. 00441.

[25] M Feng, LY Loy, F Zhang, and C Guan. Artifact removal for intracranial pressure monitoring signals: a robust solution with signal decomposition. In *Conf Proc IEEE Eng Med Biol Soc*, pages 797–801. American Medical Informatics Association, 2011.

[26] AS Fialho, LA Celi, F Cismondi, SM Vieira, SR Reti, JM Sousa, SN Finkelstein, et al. Disease-based modeling to predict fluid response in intensive care units. *Methods Inf Med*, 52(6):494–502, 2013.

[27] Writing Group for the Women's Health Initiative Investigators et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the women's health initiative randomized controlled trial. *Jama*, 288(3):321–333, 2002.

[28] Rafael Gabriel Sanchez, Luis Maria Sanchez Gomez, Loreto Carmona, Marta Roqué i Figuls, and Xavier Bonfill Cosp. Hormone replacement therapy for preventing cardiovascular disease in post-menopausal women. *The Cochrane Library*, 2005.

[29] Charles J Geyer. Handbook of Markov Chain Monte Carlo. 20116022:295–311, May 2011.

[30] Marzyeh Ghassemi, Tristan Naumann, Finale Doshi-Velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits. Unfolding physiological state: Mortality modelling in intensive care units. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 75–84. ACM, 2014.

[31] Marzyeh Ghassemi, Tristan Naumann, Rohit Joshi, and Anna Rumshisky. Topic models for mortality modeling in intensive care units. In *ICML Workshop on Machine Learning for Clinical Data Analysis*, volume Poster. International Conference on Machine Learning, 2012.

[32] Marzyeh Ghassemi, Marco AF Pimentel, Tristan Naumann, Thomas Brennan, David A Clifton, Peter Szolovits, and Mengling Feng. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in icu with sparse, heterogeneous clinical data. In *Proc. Twenty-Ninth AAAI Conf. on Artificial Intelligence*, 2015.

[33] Marzyeh Ghassemi, Zeeshan Syed, Daryush D Mehta, Jarrad H Van Stan, Robert E Hillman, and John V Guttag. Uncovering voice misuse using symbolic mismatch. In *JMLR (Journal of Machine Learning Research): MLHC Conference Proceedings*, 2016.

[34] Marzyeh Ghassemi, Jarrad H Van Stan, Daryush D Mehta, Matías Zañartu, Harold A Cheyne, Robert E Hillman, and John V Guttag. Learning to detect vocal hyperfunction from ambulatory neck-surface acceleration features: Initial results for vocal fold nodules. *IEEE Transactions on Biomedical Engineering*, 61(6):1668–1675, 2014.

[35] Marzyeh Ghassemi, Mike Wu, Mengling Feng, Leo A Celi, Peter Szolovits, and Finale Doshi-Velez. Understanding vasopressor intervention and weaning: Risk prediction in a public heterogeneous clinical time series database. *Journal of the American Medical Informatics Association*, page ocw138, 2016.

[36] Marzyeh Ghassemi, Mike Wu, Michael Hughes, and Finale Doshi-Velez. Predicting intervention onset in the icu with switching state space models. In *Proceedings of the AMIA Summit on Clinical Research Informatics (CRI)*, volume 2017. American Medical Informatics Association, 2017.

[37] T. Griffiths and M. Steyvers. Finding scientific topics. In *PNAS*, volume 101, pages 5228–5235, 2004.

[38] Neil A Halpern and Stephen M Pastores. Critical care medicine in the united states 2000–2005: an analysis of bed numbers, occupancy rates, payer mix, and costs. *Critical care medicine*, 38(1):65–71, 2010.

[39] Katharine E Henry, David N Hager, Peter J Pronovost, and Suchi Saria. A targeted real-time early warning score (trewscore) for septic shock. *Science Translational Medicine*, 7(299):299ra122–299ra122, 2015.

[40] Dave Higdon. Space and space-time modeling using process convolutions. In *Quantitative methods for current environmental issues*, pages 37–56. Springer, 2002.

[41] Roman Hlatky, Alex B Valadka, and Claudia S Robertson. Intracranial pressure response to induced hypertension: role of dynamic pressure autoregulation. *Neurosurgery*, 57(5):917–923, 2005.

[42] Ming-Wang Hsiung and Yu-Che Hsiao. The characteristic features of muscle tension dysphonia before and after surgery in benign lesions of the vocal fold. *ORL; journal for oto-rhino-laryngology and its related specialties*, 66(5):246–254, 2003.

[43] Caleb W Hug and Peter Szolovits. Icu acuity: real-time models versus daily models. In *AMIA Annual Symposium Proceedings*, volume 2009, page 260. American Medical Informatics Association, 2009.

[44] Kristel JM Janssen, A Rogier T Donders, Frank E Harrell, Yvonne Vergouwe, Qingxia Chen, Diederick E Grobbee, and Karel GM Moons. Missing covariate data in medical research: to impute is better than to ignore. *Journal of clinical epidemiology*, 63(7):721–727, 2010.

[45] Alistair EW Johnson, Andrew A Kramer, and Gari D Clifford. A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy*. *Critical care medicine*, 41(7):1711–1718, 2013.

[46] Rohit Joshi and Peter Szolovits. Prognostic physiology: modeling patient severity in intensive care units using radial domain folding. In *AMIA Annual Symposium Proceedings*, volume 2012, page 1276. American Medical Informatics Association, 2012.

[47] William A Knaus, DP Wagner, EA e a1 Draper, JE Zimmerman, Marilyn Bergner, P Gl Bastos, CA Sirio, DJ Murphy, T Lotring, and A Damiano. The apache iii prognostic system. risk prediction of hospital mortality for critically ill hospitalized adults. *CHEST Journal*, 100(6):1619–1636, 1991.

[48] Kanak Bikram Kshetri. *Modelling patient states in intensive care patients*. PhD thesis, Massachusetts Institute of Technology, 2011.

[49] Giovanni Landoni, Marco Comis, Massimiliano Conte, Gabriele Finco, Marta Mucchetti, Gianluca Paternoster, Antonio Pisano, Laura Ruggeri, Gabriele Alvaro, Manuela Angelone, et al. Mortality in multicenter critical care trials: an analysis of interventions with a significant effect. *Critical care medicine*, 43(8):1559–1568, 2015.

[50] Thomas A Lasko, Joshua C Denny, and Mia A Levy. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PloS one*, 8(6):e66341, 2013.

[51] John Laver. The phonetic description of voice quality. *Cambridge Studies in Linguistics London*, 31:1–186, 1980.

[52] J.R. Le Gall, S. Lemeshow, and F. Saulnier. A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *JAMA*, 270(24):2957–2963, 1993.

[53] Cheng H Lee, Natalia M Arzeno, Jonathan C Ho, Haris Vikalo, and Joydeb Ghosh. An imputation-enhanced algorithm for icu mortality prediction. In *Computing in Cardiology (CinC), 2012*, pages 253–256. IEEE, 2012.

[54] Joon Lee and Roger G Mark. An investigation of patterns in hemodynamic data indicative of impending hypotension in intensive care. *Biomedical engineering online*, 9(1):62, 2010.

[55] Li-wei Lehman, Mohammed Saeed, William Long, Joon Lee, and Roger Mark. Risk stratification of icu patients using topic models inferred from unstructured progress notes. In *AMIA Annual Symposium Proceedings*, volume 2012, page 505. American Medical Informatics Association, 2012.

[56] LW Lehman, RP Adams, L Mayaud, GB Moody, A Malhotra, RG Mark, and S Nemati. A physiological time series dynamics-based approach to patient monitoring and outcome prediction. *IEEE journal of biomedical and health informatics*, 19(3):1068, 2015.

[57] Q Li and GD Clifford. Dynamic time warping and machine learning for signal quality assessment of pulsatile signals. *Physiological Measurement*, 33(9):1491, 2012.

[58] Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Bill Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 2–11. ACM, 2003.

[59] Jessica Lin, Eamonn Keogh, Li Wei, and Stefano Lonardi. Experiencing sax: a novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, 15(2):107–144, 2007.

[60] Peng Liu, Lei Lei, Junjie Yin, Wei Zhang, Wu Naijun, and Elia El-Darzi. Healthcare data mining: predicting inpatient length of stay. 2006.

[61] Benjamin M Marlin, David C Kale, Robinder G Khemani, and Randall C Wetzel. Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 389–398. ACM, 2012.

[62] Andrea Marshall, Douglas G Altman, Patrick Royston, and Roger L Holder. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC medical research methodology*, 10(1):7, 2010.

[63] MATLAB. *Signal Processing Toolbox Release 2013b*. The MathWorks, Inc.

[64] J Michael McGinnis, Leigh Stuckhardt, Robert Saunders, Mark Smith, et al. *Best care at lower cost: the path to continuously learning health care in America*. National Academies Press, 2013.

[65] Daryush D Mehta, Matias Zanartu, Shengran W Feng, Harold A Cheyne, and Robert E Hillman. Mobile voice health monitoring using a wearable accelerometer sensor and a smartphone platform. *Biomedical Engineering, IEEE Transactions on*, 59(11):3090–3096, 2012.

[66] Arman Melkumyan and Fabio Ramos. Multi-kernel gaussian processes. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, IJCAI'11, pages 1408–1413. AAAI Press, 2011.

[67] Edward J Mills, Kristian Thorlund, and John PA Ioannidis. Demystifying trial networks and network meta-analysis. *Bmj*, 346:f2914, 2013.

[68] Marcus Müllner, Bernhard Urbanek, Christof Havel, Heidrun Losert, Gunnar Gamper, and Harald Herkner. Vasopressors for shock. *The Cochrane Library*, 2004.

[69] OECD. Oecd labour force statistics 2014.

[70] Office of the National Coordinator for Health Information Technology. Office-based physician electronic health record adoption. *Health IT Quick-Stat 50*, 2016.

[71] Ann-Christine Ohlsson, Olle Brink, and Anders Lofqvist. A voice accumulatorâĂŤvalidation and application. *Journal of Speech, Language, and Hearing Research*, 32(2):451–457, 1989.

[72] World Health Organization. Noncommunicable diseases-fact sheet, 2015. 2015.

[73] Gustavo A Ospina-Tascón, Gustavo Luiz Büchele, and Jean-Louis Vincent. Multicenter, randomized, controlled trials evaluating mortality in intensive care: Doomed to fail? *Critical care medicine*, 36(4):1311–1322, 2008.

[74] Pranav Patel, Eamonn Keogh, Jessica Lin, and Stefano Lonardi. Mining motifs in massive time series databases. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pages 370–377. IEEE, 2002.

[75] Adler Perotte, Rajesh Ranganath, Jamie S Hirsch, David Blei, and Noémie Elhadad. Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. *Journal of the American Medical Informatics Association*, 22(4):872–880, 2015.

[76] Maiyaporn Phanich, Phathrajarin Pholkul, and Suphakant Phimoltares. Food recommendation system using clustering analysis for diabetic patients. In *Information Science and Applications (ICISA), 2010 International Conference on*, pages 1–8. IEEE, 2010.

[77] Gianluigi Pillonetto, Francesco Dinuzzo, and Giuseppe De Nicolao. Bayesian online multitask learning of gaussian processes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(2):193âĂŞ205, 2010. 00025.

[78] Ross L Prentice, Robert D Langer, Marcia L Stefanick, Barbara V Howard, Mary Pettinger, Garnet L Anderson, David Barad, J David Curb, Jane Kotchen, Lewis Kuller, et al. Combined analysis of women's health initiative observational and clinical trial data on postmenopausal hormone treatment and cardiovascular disease. *American Journal of Epidemiology*, 163(7):589–599, 2006.

[79] John A Quinn, Christopher KI Williams, and Neil McIntosh. Factorial switching linear dynamical systems applied to physiological condition monitoring. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(9):1537–1551, 2009.

[80] Joaquin QuiÃśonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *J. Mach. Learn. Res.*, 6:1939âĂŞ1959, December 2005.

[81] Lawrence R Rabiner and Biing-Hwang Juang. An introduction to hidden Markov models. *ASSP Magazine, IEEE*, 3(1):4–16, 1986.

[82] Leena Rantala and Erkki Vilkman. Relationship between subjective voice complaints and acoustic parameters in female teachers' voices. *Journal of Voice*, 13(4):484–495, 1999.

[83] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2006.

[84] Nelson Roy, Julie Barkmeier-Kraemer, Tanya Eadie, M Preeti Sivasankar, Daryush Mehta, Diane Paul, and Robert Hillman. Evidence-based clinical voice assessment: A systematic review. *American Journal of Speech-Language Pathology*, 22(2):212–226, 2013.

[85] Nelson Roy, Ray M Merrill, Steven D Gray, and Elaine M Smith. Voice disorders in the general population: prevalence, risk factors, and occupational impact. *The Laryngoscope*, 115(11):1988–1995, 2005.

[86] A Rumshisky, M Ghassemi, T Naumann, P Szolovits, VM Castro, TH McCoy, and RH Perlis. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Translational Psychiatry*, 6(10):e921, 2016.

[87] M. Saeed et al. Multiparameter Intelligent Monitoring in Intensive Care II: A public-access intensive care unit database. *Critical Care Medicine*, 39(5):952–960, May 2011.

[88] Mohammed Saeed and Roger Mark. A novel method for the efficient retrieval of similar multiparameter physiologic time series using wavelet-based symbolic representations. In *AMIA Annual Symposium Proceedings*, volume 2006, page 679. American Medical Informatics Association, 2006.

[89] Mohammed Saeed, Mauricio Villarroel, Andrew T Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H Kyaw, Benjamin Moody, and Roger G Mark. Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access intensive care unit database. *Critical care medicine*, 39(5):952, 2011.

[90] G. Salton and C. S. Yang. On the specification of term values in automatic indexing. *Journal of Documentation*, 29(4):351–372, 1973.

[91] Suchi Saria, Daphne L Koller, and Anna A Penn. Learning individual and population level traits from clinical temporal data. In *In the Predictive Models in Personalized Medicine Workshop*. Twenty-Fourth Annual Conference on Neural Information Processing Systems, 2012.

[92] Suchi Saria, Gayle McElvain, Anand K Rajani, Anna A Penn, and Daphne L Koller. Combining structured and free-text data for automatic coding of patient outcomes. In *AMIA Annual Symposium Proceedings*, volume 2010, page 712. American Medical Informatics Association, 2010.

[93] George CM Siontis, Ioanna Tzoulaki, and John PA Ioannidis. Predicting death: an empirical evaluation of predictive tools for mortality. *Archives of internal medicine*, 171(19):1721–1726, 2011.

[94] Sheila V Stager, Rebecca Neubert, Susan Miller, Joan Roddy Regnell, and Steven A Bielamowicz. Incidence of supraglottic activity in males and females: a preliminary report. *Journal of Voice*, 17(3):395–402, 2003.

[95] O. Stegle, S.V. Fallert, D. J C MacKay, and S. Brage. Gaussian process robust regression for noisy heart rate data. *Biomedical Engineering, IEEE Transactions on*, 55(9):2143–2151, Sept 2008.

[96] Oliver Stegle, Christoph Lippert, Joris M. Mooij, Neil D. Lawrence, and Karsten M. Borgwardt. Efficient inference in matrix-variate gaussian models with$\backslash$ iid observation noise. In *Advances in Neural Information Processing Systems*, page 630âĂŞ638, 2011.

[97] Fahim Sufi, Ibrahim Khalil, and Zahir Tari. A cardiod based technique to identify cardiovascular diseases using mobile phones and body sensors. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 5500–5503. IEEE, 2010.

[98] Kevin Swersky, Jasper Snoek, and Ryan P Adams. Multi-task bayesian optimization. In *Advances in Neural Information Processing Systems*, pages 2004–2012, 2013.

[99] Zeeshan Syed, John Guttag, and Collin Stultz. Clustering and symbolic analysis of cardiovascular signals: discovery and visualization of medically relevant patterns in long-term data using limited prior knowledge. *EURASIP Journal on Applied Signal Processing*, 2007(1):97–97, 2007.

[100] Zeeshan Syed and John V Guttag. Unsupervised similarity-based risk stratification for cardiovascular events using long-term time-series data. *Journal of Machine Learning Research*, 12:999–1024, 2011.

[101] Martin J Tobin. Principles and practice of mechanical ventilation, 2006.

[102] J-L Vincent, Rui Moreno, Jukka Takala, S Willatts, Arnaldo De Mendonça, Hajo Bruining, CK Reinhart, PM Suter, and LG Thijs. The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. *Intensive care medicine*, 22(7):707–710, 1996.

[103] Jean-Louis Vincent. Critical care-where have we been and where are we going? *Critical Care*, 17(Suppl 1):S2, 2013.

[104] Jean-Louis Vincent, John C Marshall, Silvio A Ñamendys-Silva, Bruno FranÇois, Ignacio Martin-Loeches, Jeffrey Lipman, Konrad Reinhart, Massimo Antonelli, Peter Pickkers, Hassane Njimi, et al. Assessment of the worldwide burden of critical illness: the intensive care over nations (icon) audit. *The lancet Respiratory medicine*, 2(5):380–386, 2014.

[105] Jean-Louis Vincent and Mervyn Singer. Critical care: advances and future perspectives. *The Lancet*, 376(9749):1354–1361, 2010.

[106] Hans Wackernagel. *Multivariate Geostatistics.* Springer, April 2003.

[107] Xiaoyue Wang, Abdullah Mueen, Hui Ding, Goce Trajcevski, Peter Scheuermann, and Eamonn Keogh. Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, 26(2):275–309, 2013.

[108] C Werner and K Engelhard. Pathophysiology of traumatic brain injury. *Br. J. Anaesth*, 99(1):4–9, 2007.

[109] Jenna Wiens, Eric Horvitz, and John V Guttag. Patient risk stratification for hospital-associated c. diff as a time-series classification task. In *Advances in Neural Information Processing Systems*, pages 476–484, 2012.

[110] Karl L Yang and Martin J Tobin. A prospective study of indexes predicting the outcome of trials of weaning from mechanical ventilation. *New England Journal of Medicine*, 324.

Appendix A

Appendices

A.1 Additional Information for Chapter 2

A.1.1 Extracting and Processing Signals

A total of 24 variables corresponding to vital signs, lab results, and static demographic data were extracted from the MIMIC II 2.6 database. Data was gathered from four ICUs at the Beth Israel Deaconess Medical Center (BIDMC): medical (MICU), surgical (SICU), coronary care unit (CCU), and cardiac surgery recovery unit (CSRU).

A patient's 30-day mortality was extracted by subtracting the patient's date of death from time of release from ICU. If the difference was less than 30 days, the patient was removed from consideration. Physiological variables were the timeseries of 6 nurse-verified vital signs heart-rate (HR), mean arterial blood pressure (MeanBP), blood oxygenation level (SPO2), temperature (TEMP), spontaneous respiration rate (RESP), and urine output (URINE); the timeseries of 11 laboratory measurements blood urea nitrogen (BUN), hematocrit (HCT), creatinine (CREAT), bicarbonate (BICAR), lactate (LACT), magnesium (Mg), potassium (K), sodium (Na), glucose (GLU), platelet count (PC), and white blood cell count (WBC); and 7 static variables admitting age, gender, first SAPS I score, first SOFA acuity score, first weight, use of pacemaker, and whether the patient was noted as ”at risk” for falls. Timeseries variables were first binned into hours from when the patient was admitted, and the value for that hour was the mean for that hour. To handle missing data, we only incorporated features with greater than 10% non-missing entries (MEAN BP, TEMP, HR, SPO2, FIO2, RR, GLU, BICAR, HCT, K), and smoothed the data through sample-and-hold.

Extracting and Processing Outcomes We extracted vassopressor administration as any medication event with a generic or brand-name vassopressor label, including dopamine, epinephrine, isuprel, levophed, vasopressin, and neosynephrine. We considered any modification of vassopressor settings to be a binary indicator of vassopressor administration in the hour it occurred in. Because continuing vassopressor administration is not always noted in the electronic health record, we interpolated any vassopressor gaps less than 4 hours as being continuously on the medication, unless there was an explicit stoppage of the medication noted.

From this smoothed timeseries, we computed the start time of the i^{th} vassopressor administration $t_n^{v_i}$ and its corresponding wean $t_n^{w_i}$ for each patient n. Table 2.1 compares the values of individual variables for patients who received vassopressors to patients who did not (the controls C). For patients who did receive a vassopressor, we denote the time until the first administration $t = 1 \dots t^{v_1}$ as V- and the time between the first administration and the first wean $t = t^{v_1} + 1 \dots t^{w_1}$ as V+.

A.2 Additional Information for Chapter 3

Patient Cohort Sizes

Table A.1.

A.2.1 List of Inferred Topics

Table A.2.

Topic	Top Ten Words
1	cabg, pain, ct, artery, coronary, valve, post, wires, chest, sp

2	ccu, cath, mg, am, sp, groin, bp, cardiac, hr, cont
3	picc, line, name, procedure, catheter, vein, tip, placement, clip, access
4	biliary, mass, duct, metastatic, bile, cancer, left, ca, tumor, clip
5	liver, renal, hepatic, ascites, dialysis, failure, flow, transplant, portal, ultrasound
6	ct, contrast, pelvis, abdomen, fluid, bowel, clip, free, wcontrast, iv
7	thick, secretions, vent, trach, resp, tf, tube, coarse, cont, suctioned
8	chest, pneumothorax, tube, reason, clip, sp, ap, left, portable, ptx
9	remains, family, gtt, line, map, cont, levophed, cvp, bp, levo
10	name, neo, gtt, stitle, dr, sbp, resp, cont, wean, aware
11	remains, increased, temp, hr, pt, cc, ativan, cont, mg, continues
12	micu, code, stool, hr, bp, social, note, id, received, cchr
13	chest, pulmonary, bilateral, edema, portable, clip, reason, ap, pleural, effusions
14	resp, cough, sats, mask, sob, wheezes, nc, status, mg, neb
15	intubated, vent, ett, secretions, propofol, abg, respiratory, resp, care, sedated
16	gtt, insulin, bs, lasix, endo, monitor, mg, am, plan, iv
17	drainage, pain, abd, fluid, draining, drain, incision, sp, intact, pt
18	heparin, afib, ptt, am, gtt, mg, rate, hr, pvcs, iv
19	name, pacer, namepattern, placement, heart, pacemaker, ventricular, av, rate, chest
20	left, lung, effusion, lobe, pleural, lower, chest, upper, ct, opacity
21	skin, noted, care, left, applied, changed, draining, coccyx, wound, edema
22	tube, placement, tip, line, portable, ap, reason, position, chest, ng
23	noted, shift, name, pt, patent, patient, foley, agitated, soft, mg
24	hct, pt, gi, blood, bleeding, am, stable, unit, bleed, noted

25	name, am, mg, able, bp, time, night, times, doctor, confused
26	pain, co, denies, oriented, neuro, plan, diet, po, pt, floor
27	name, family, neuro, care, noted, status, plan, stitle, dr, remains
28	clip, reason, ro, medical, examination, evidence, impression, underlying, condition, normal
29	neuro, sbp, bp, commands, iv, cough, soft, status, lopressor, swallow
30	skin, stable, social, family, intact, tsicu, id, note, support, endo
31	woman, female, husband, name, pain, patient, pm, am, hospital, noted
32	diagnosis, admitting, name, reason, please, examination, yearold, eval, findings, underlying
33	name, neck, soft, patient, noted, anterior, epidural, level, posterior, namepattern
34	ct, contrast, chest, lymph, optiray, images, lesions, iv, nodes, lobe
35	left, stenosis, disease, clip, reason, carotid, severe, report, radiology, final
36	femoral, foot, left, leg, iliac, groin, lower, patent, graft, extremity
37	acute, reason, head, clip, evidence, eval, name, wo, status, ct
38	aortic, aorta, cta, wwo, dissection, recons, contrast, left, aneurysm, chest
39	left, ivc, filter, vein, pulmonary, veins, dvt, clip, inferior, upper
40	left, fracture, ap, views, reason, clip, hip, distal, lat, report
41	spine, cervical, spinal, clip, thoracic, fall, lumbar, vertebral, contrast, reason
42	hemorrhage, head, ct, left, frontal, contrast, subdural, hematoma, clip, bleed
43	ct, trauma, contrast, injury, fracture, fractures, pelvis, clip, wcontrast, sp
44	contrast, brain, head, left, mri, images, mra, stroke, clip, cerebral

45	catheter, name, procedure, contrast, wire, french, placed, needle, advanced, clip
46	artery, left, common, distal, catheter, internal, branches, flow, name, middle
47	vein, stent, catheter, name, mm, portal, tips, balloon, venous, sheath
48	service, distinct, procedural, artery, sel, carotid, left, cath, name, clip
49	catheter, name, performed, embolization, contrast, bleeding, procedure, mesenteric, extravasation, clip
50	artery, carotid, left, aneurysm, injection, vertebral, internal, evidence, clip, cerebral

Table A.2: Top ten most probable words for all topics.

Table A.1: Patient cohort size at each time tested by time-varying models. Note that patients are removed from a prediction time if they are discharged or die prior to that time.

		Cohort Size (Control, Positive)		
Time (Hours)	**Total**	**In-Hospital**	**30 Day**	**1 Year**
0	5784	5157, 627	5597, 187	5058, 726
12	5784	5157, 627	5597, 187	5058, 726
24	5749	5128, 621	5563, 186	5026, 723
36	5563	4998, 565	5382, 181	4855, 708
48	5497	4937, 560	5318, 179	4795, 702
60	5161	4664, 497	4986, 175	4480, 681
72	5084	4591, 493	4911, 173	4407, 677
84	4691	4241, 450	4524, 167	4043, 648
96	4587	4140, 447	4421, 166	3945, 642
108	4116	3710, 406	3963, 153	3530, 586
120	4030	3626, 404	3877, 153	3448, 582
132	3570	3210, 360	3427, 143	3023, 547
144	3496	3141, 355	3354, 142	2956, 540
156	3026	2707, 319	2898, 128	2533, 493
168	2967	2652, 315	2840, 127	2479, 488
180	2580	2291, 289	2468, 112	2138, 442
192	2541	2254, 287	2431, 110	2109, 432
204	2215	1953, 262	2117, 98	1825, 390
216	2186	1925, 261	2090, 96	1802, 384
228	1925	1681, 244	1837, 88	1575, 350