Time Series Analysis of Nursing Notes for Mortality Prediction via a State Transition Topic Model

Yohan Jo
Carnegie Mellon University
yohanj@cs.cmu.edu

Natasha Loghmanpour
Carnegie Mellon University
nloghman@cmu.edu

Carolyn Penstein Rosé
Carnegie Mellon University
cprose@cs.cmu.edu

ABSTRACT

Accurate mortality prediction is an important task in intensive care units in order to channel prompt care to patients in the most critical condition and to reduce nurses' alarm fatigue. Nursing notes carry valuable information in this regard, but nothing has been reported about the effectiveness of temporal analysis of nursing notes in mortality prediction tasks.

We propose a time series model that uncovers the temporal dynamics of patients' underlying states from nursing notes. The effectiveness of this information in mortality prediction is examined for mortality prediction for five different time spans ranging from one day to one year. Our experiments show that the model captures both patient states and their temporal dynamics that have a strong correlation with patient mortality. The results also show that incorporating temporal information improves performance in long-term mortality prediction, but has no significant effect in short-term prediction.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life and Medical Sciences—*Medical information system*; I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems—*Medicine and Science*; H.2.8 [**Database Management**]: Database Applications—*Data mining*; G.3 [**Mathematics of Computing**]: Probability and Statistics—*Time series analysis*

General Terms

Algorithms, Human Factors, Languages

Keywords

Healthcare, Medical Data Mining, Nursing Notes, Mortality Prediction, Hidden Markov Model, Latent Dirichlet Allocation, State Transition Topic Model

1. INTRODUCTION

Predicting a patient's risk of mortality and taking appropriate action are important activities in intensive care units (ICUs). High accuracy in mortality prediction helps nurses manage patient care by placing patients in different priority queues. It also enhances nurses' efficiency by reducing the number of false alarms, which cause them alarm fatigue and desensitize them to real alarms [6]. In this respect, nursing notes contain valuable information to inform more accurate prediction models. This information includes nurses' observations and intuitions that do not fit into the accompanying recorded structured data. Nursing notes have the potential to uncover hidden clues about a patient's health and mental state as they change over time, such as the factors of family support and mental fitness (Figure 1).

PROB: S/P AVR FUNCTIONAL HEALTH PATTERNS AND HISTORY COMPLETED IN CHART.
NEURO: PT AWAKE, FOLLOWING COMMANDS. MAE, NODS APPROPRIATELY. PERL.
CV: SR NO VEA NOTED. CONT ON MILRINONE AND NEO. CO/CI [**7-26**]. K REPLACED MULTIPLE TIMES. CT DRAINING S/S DRAINAGE. PACER OFF. MORPHINE FOR PAIN X3 WITH GOOD EFFECT.
RESP: PT WEANING. LUNGS WITH WHEEZES, CLEARED WITH COMBIVENT INHALER. SUCTION FOR THICK CLEAR/YELLOW SPUTUM. PT STILL ACIDOTIC, IMV 16, TV 700.
GI: NGT TO LOW CONT SUCTION, NO DRAINAGE,
ENDO: INITIAL BS ELEVATED AND TREATED PER PROTOCOL.
GU: ADEQUATE AMOUNT OF CLEAR YELLOW URINE. PLACEMENT CHECKED-GOOD.
SOCIAL: FAMILY HERE TO VISIT.
ASSESSMENT: WEANING SLOWLY
PLAN: RECHECK ABGS, LYTES. SUCTION PRN. MED FOR PAIN. CONT VENT WEAN.

Figure 1: Example Nursing Note

However, the potential of nursing notes in informing mortality prediction has started to be investigated only recently [8, 14]. Furthermore, there is no previous work on temporal analysis of nursing notes and its use for mortality prediction, while time series analysis of patient signals and clinical data has shown a lot of benefits in patient outcome prediction [5, 23, 13, 18]. Hence, this paper offers the following contributions to this important research field.

- Proposes and evaluates a model to uncover the temporal dynamics of underlying patient states from nursing notes.
- Evaluates the effectiveness of the identified temporal dynamics for improving mortality prediction.
- Offers qualitative insight into different types of textual features regarding their roles in mortality prediction.

We propose a novel and intuitive model that combines a hidden Markov model (HMM) and latent Dirichlet allocation (LDA) [4]. This model assumes that there are a set of underlying patient states, and every pair of states has a meaningful transition probability. Each state is represented by a topic distribution, from which nursing notes are generated. The model, when applied to nursing notes, learns topics embedded in nursing notes, the topic distribution of each state and each nursing note, and state transition prob-

CIKM'15, October 19–23, 2015, Melbourne, Australia.

DOI: http://dx.doi.org/10.1145/2806416.2806541.

abilities. We can also estimate a sequence of states each patient is identified as having transitioned through, which is then used for mortality prediction. This model is general enough to be applied for the purpose of identifying temporal dynamics in any text stream. Our simple and general model shows effectiveness in uncovering latent patient states in nursing notes and improving mortality prediction. The model is described in detail in Section 4.

Three tasks are evaluated in order to validate the learned temporal information and its benefit for mortality prediction. In Task 1 (Section 6.1), we qualitatively describe the learned topics, underlying patient states, and state transition patterns revealed from ICU nursing notes. In Task 2 (Section 6.2), we perform mortality prediction and show that including the temporal information achieves a significant improvement for long-term predictions, but little effect in short-term prediction. In Task 3 (Section 6.3), we examine individual textual features more in detail. Along with the temporal information, we investigate two other types of textual features—n-grams and standard topics—with respect to their roles and limitations in mortality prediction. To the best of our knowledge, this article is the most comprehensive treatment of the use of textual features for mortality prediction to date.

We first introduce related work (Section 2) and formally define the problem (Section 3). Next, we explain our model in detail (Section 4). For experiments, we first describe the experimental settings (Section 5), and perform the three tasks aforementioned (Section 6). Finally, we conclude the paper by summarizing the findings and discuss limitations (Section 7).

2. RELATED WORK

This section describes previous work on *patient mortality prediction based on free-text medical documents* and *joint models of topics and time*. Only recently did free-text medical documents—nursing notes, discharge summaries, laboratory test reports, radiology reports, etc.— start to be employed for mortality prediction [14, 8]. These studies used medical notes from MIMIC II Clinical Database and employed topic modeling approaches. Specifically, Lehman et al. [14] extracted topics from nursing notes using the hierarchical Dirichlet process [19] and used the learned topic distributions as input to logistic linear regression for predicting each patient's mortality. Ghassemi et al. [8], whose work is the most related work with ours, took a similar approach, except that they learned topics using latent Dirichlet allocation and used a support vector machine for mortality prediction. These studies showed the promise of nursing notes, with which they achieved higher accuracy in mortality prediction than with admission-time patient information such as ages and Simplified Acute Physiology Scores. Our work adopts these approaches as baselines.

This paper presents the first work to model the temporal dynamics of patients' states from nursing notes and apply this information to mortality prediction. The work by Ghassemi et al. [8] and our work are different in several aspects. Their work used nursing notes, laboratory test results, and radiology reports, but our work focuses only on nursing notes. Also, they excluded NICU (neonatal ICU) patients, but we include all types of patients. This actually reveals an interesting pattern of NICU patients in their state transitions. Their work made prediction for in-hospital and post-discharge mortality, but our work predicts 1-day, 1-week, 1-month, 6-month, and 1-year mortality, which may be of more interest to patients, families, and nurses and doctors for taking appropriate action.

There have been a lot of efforts to incorporate temporal aspects into topic models for other types of text. One big category of prior approaches attempts to model topic evolution over time [2, 7, 11, 15]. Under the assumption that the language models of topics change over time, these models discover topic dynamics from time-stamped documents. Our work is not aiming at modeling topic evolution. Some other models instead assume that topic popularity changes over time [10, 21]. In these models, the probability of a topic being manifested depends on the timestamp of the document. This trend, however, is too general to apply to different conditions of individual patients, and these models do not assume any underlying states of patients. There is also a set of segment-level time-aware topic models [3, 9, 20, 26]. These models suppose that the topic of a segment or word is affected by the topic of the previous segment or word within the document. This assumption is effective in modeling topic consistency within a segment of text and topic transition between segments. However, this topic transition can hardly represent the transition of patient states.

Not as popular as the previous categories, a few topic models consider the inter-dependency of topic distributions between documents. These models can again be categorized into two subgroups. In the first subgroup, the topic distribution of a document directly determines the next document's topic distribution usually via a transition matrix [17, 22, 25]. This assumption may be too strong for nursing notes since the topics of a nursing note are dependent on the complex details of a patients' condition, rather than being unambiguously determined by the previous note's topic distribution. Indeed, the mixed membership Markov model [17] did not produce interpretable state sequences and performed poorly in mortality prediction.

The second subgroup of models, to which our model belongs, assumes hidden underlying states, and each state is associated with a topic distribution from which documents are generated [24]. The state of a document is probabilistically determined by the previous document's state. In our task, this type of model is reasonable in that a patient's state is probabilistically determined by the previous state, and each state generates documents with certain topics. Also, the finite set of underlying states, unlike the infinite state space of the first subgroup, makes it straightforward to draw a state transition diagram such as Figure 4. Although the other model in this group [24] has a similar generative process, it assumes a set of independent state sequences, and each sequence has no branch factors. That is, slightly different state paths can be made only by independent sets of states and state sequences, which makes it nontrivial to compare the states of different patients. This is the main difference from our model, which assumes a shared set of states and branch factors.

3. MORTALITY PREDICTION

We define mortality prediction as the prediction of a patient's mortality or survival from a reference time within a given timeframe (e.g., one day, one month, etc.) on the basis of the patient's nursing notes up to the reference time. The prediction task will be used to validate our model (Section 4), which gives visibility to sequences of patients' latent

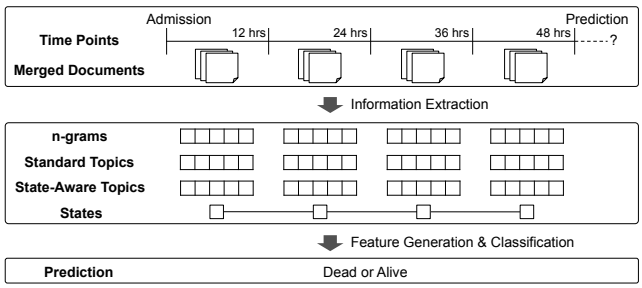


Figure 2: Process of Mortality Prediction

states through the interpretation of the topic distributions of their nursing notes. The improvement over baseline in the prediction task using this representation validates that the patterns identified from nursing notes say something important about the experience of patients over time.

We perform predictions in the short-term (one day and one week) and in the long-term (one month, six months, and one year). Figure 2 illustrates the overall prediction process. For an ICU stay of a patient, we consider every 12 hours as one time point. Since some notes are too short or focus only on a single topic, all nursing notes in the same time point are merged into one document for analysis so that the merged document can reflect the overall topics at that time point. In this way, one sequence of composite documents is obtained for each patient. The task then is, given a sequence of documents of a patient, to predict the patient's mortality or survival within a given term from the time of the last document in the series, e.g., one week. One classifier is made for each time point. The classifier of the i-th time point makes prediction based on the nursing notes written until the i-th time point. The i-th classifier is trained on the nursing notes of all patients written during their first i time points. In training, the document sequences of the patients who died within the given term are counted as positive instances, and the other sequences negative instances.

4. STATE TRANSITION TOPIC MODEL

To extract the temporal dynamics of patients' underlying states from nursing notes, we propose an intuitive model called the State Transition Topic Model (STTM). STTM is a joint model integrating a hidden Markov model (HMM) and latent Dirichlet allocation (LDA). Different from the standard LDA, our model assumes latent underlying states and transition probabilities between them. Each state is represented by a probability distribution over topics from which a document is generated. STTM thus models hidden states and topics simultaneously. Consider the following scenario. A patient moves between states as time elapses according to the Markovian assumption, i.e., following state transition probability distributions. When a patient enters a state, a document is generated from the state's topic distribution according to the LDA assumption. When applied on a set of sequences of documents, STTM can learn a set of topics that constitute the documents, the topic distribution of each state and each document, and state transition probabilities. Based on this, we can estimate a patient's state at each time point, producing a sequence of states for the entire ICU stay. Note that this model is general enough to be applied not only to nursing notes, but any streams of data

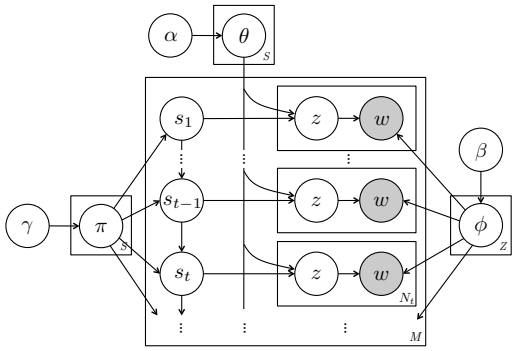


Figure 3: Graphical Representation of STTM.

Table 1: Meanings of Notations

d_t^m	the t-th document in sequence m
$w_{t,i}^m$	the i-th word in d_t^m
$z_{t,i}^m$	the topic assigned to $w_{t,i}^m$
s_t^m	the state of d_t^m
$\mathbf{w}_t^m$	all words in d_t^m
$\mathbf{z}_{t,-i}^m$	the topics of all words in d_t^m except for $w_{t,i}^m$
$\mathbf{s}_{-(m,t)}$	the states of all documents except for d_t^m
$N_{m,t,j}^{MTZ}$	the number of words in d_t^m assigned topic j
$N_{j,w}^{ZW}$	the number of words w assigned topic j
$N_{c,j}^{SZ}$	the number of words assigned topic j in state c
$N_{c,c'}^{SS}$	the number of states c followed by state c'

points. In the following sections, we formally describe the model and an inference method based on Gibbs sampling.

4.1 Formal Definition

Suppose that there are S states, Z topics, and W unique words. The probability distribution over words for topic j is denoted by ϕ_j, which is a W-dimensional vector, and the probability distribution over topics at state c is denoted by θ_c, which is a Z-dimensional vector. The transition probability distribution of state c is denoted by π_c, which is an S-dimensional vector. We assume that all initial states come from the "0th" state, a special state that has no incoming edges. Thus, $\pi_{0,c}$ represents the probability of state c being an initial state. We take the full Bayesian approach, assuming Dirichlet priors γ, α, and β for π_c, θ_c, and ϕ_j, respectively. The graphical representation is shown in Figure 3, and the meanings of notations are listed in Table 1. The generative process is as follows.

1. For each state $c = 0, ..., S$,
   (a) Draw a transition distribution $\pi_c \sim \text{Dirichlet}(\gamma)$
   (b) Draw a topic distribution $\theta_c \sim \text{Dirichlet}(\alpha)$
2. For each topic $j = 1, ..., Z$,
   (a) Draw a word distribution $\phi_j \sim \text{Dirichlet}(\beta)$
3. For each document at time point $t = 1, 2, ...$,
   (a) Choose a state $s_t \sim \text{Categorical}(\pi_{s_{t-1}})$
   (b) For each word,
      i. Choose a topic $z \sim \text{Categorical}(\theta_{s_t})$
      ii. Choose a word $w \sim \text{Categorical}(\phi_z)$

4.2 Inference

Gibbs sampling is used for inference. Each iteration samples $z^m_{t,i}$ and s^m_t according to the following conditional probabilities.

$$p(z^m_{t,i} = j|\mathbf{z}^m_{t,-i}, \mathbf{w}^m_t, s^m_t) \propto (N^{MTZ}_{m,t,j} + \alpha)\frac{N^{ZW}_{j,w^m_{t,i}} + \beta}{\sum_j (N^{ZW}_{j,w^m_{t,i}} + \beta)},$$

$$p(s^m_t = c|\mathbf{s}_{-(m,t)}, \mathbf{z}^m_t) \propto$$

$$\left(\prod_j \frac{\Gamma\left(N^{SZ}_{c,j} + \alpha + N^{MTZ}_{m,t,j}\right)}{\Gamma\left(N^{SZ}_{c,j} + \alpha\right)}\right) \frac{\Gamma\left(\sum_{j'}(N^{SZ}_{c,j'} + \alpha)\right)}{\Gamma\left(\sum_{j'}(N^{SZ}_{c,j'} + \alpha) + |\mathbf{z}^m_t|\right)} \times$$

$$(N^{SS}_{s^m_{t-1},c} + \gamma)\frac{N^{SS}_{c,s^m_{t+1}} + \mathbf{1}(s^m_{t-1} = c = s^m_{t+1}) + \gamma}{\sum_{s^m_{t+1}}\left(N^{SS}_{c,s^m_{t+1}} + \mathbf{1}(s^m_{t-1} = c = s^m_{t+1}) + \gamma\right)}.$$

Based on the sampling results, we can estimate ϕ_j, θ_c, and π_c as follows.

$$\phi_{j,w} = \frac{N^{ZW}_{j,w} + \beta}{\sum_w (N^{ZW}_{j,w} + \beta)}, \theta_{c,j} = \frac{N^{SZ}_{c,j} + \alpha}{\sum_j (N^{SZ}_{c,j} + \alpha)},$$

$$\pi_{c,c'} = \frac{N^{SS}_{c,c'} + \gamma}{\sum_{c'} (N^{SS}_{c,c'} + \gamma)}, \theta^m_{t,j} = \frac{N^{MTZ}_{c,c'} + \alpha}{\sum_{c'} (N^{MTZ}_{c,c'} + \alpha)},$$

where θ^m_t is the document-wise topic distribution of d^m_t. A more detailed derivation process and the source code are available on our website[1].

5. EXPERIMENTAL SETTINGS

This section describes our experimental settings in detail.

5.1 Data

We use MIMIC II Clinical Database[2], which contains comprehensive clinical data of ICU patients collected between 2001 and 2008. This database contains information such as the demographics of de-identified patients, the records of their ICU and hospital admissions, nursing notes and reports, laboratory test results, and medications. Among this information, we use nursing notes for mortality prediction (Figure 1). The content usually includes the patient's neurological state, laboratory test results, medications, descriptions of facial expressions, social activities, and the nurse's impression and plans. We exclude radiology reports to make clear the impact of nursing notes, which are yet to be fully investigated for mortality prediction; however, in a real world application of the technology, those reports could be used as well to enhance the performance. Radiology reports are different from nursing notes in that they are grounded results of radiology tests but exist only for certain types of patients. Like the study by Ghassemi et al. [8], discharge summaries are also excluded since they include patient outcomes, though again, in a real world application these could be included for long term predictions. The purpose of our evaluation here is to evaluate one specific source of prediction, not to achieve the highest possible performance that could be achieved by including all available indicators.

Mortality prediction is meant to be performed during the first ICU stays of patients. We thus retrieve all nursing notes of patients written in their first ICU stay. Nursing notes are constrained to be longer than 100 characters excluding spaces (shorter notes are discarded). Nursing notes from the same time point (12 hour segment) are merged into a single document. Table 2(a) shows the statistics of the generated sequences, and Table 2(b) shows the characteristics of patients by different ICU types. The list of subject_id's is available on our website.

Table 2: Data Statistics

(a) Statistics of retrieved nursing notes

# of sequences (=# of patients)	8,808
# of notes	187,808
# of merged documents	97,769
Avg length of merged documents	1,548 chars (Stddev=904)
Avg length of sequences	11 (Stddev=23)

(b) Characteristics of ICU types

	NICU	*CSRU*	*MICU*	*CCU*	*FICU*	*SICU*
# of patients	2332	2244	1918	1358	711	245
Avg. stay	10	4	5	4	5	3

(c) Percentages of patients who died within certain periods of time after admission

Died within	*1 day*	*1 week*	*1 month*	*6 months*	*1 year*
%	0.03	0.08	0.14	0.19	0.22

Nursing notes are preprocessed as follows. First, in order to reduce the sparsity of de-identified information, the de-identified pieces of text, e.g., patient names, are normalized to their category name. For example, "[**Female First Name (un) 978**]" is normalized to "PHI_FEMALEFIRSTNAME". Second, to distinguish whether a situation is positive or negative, simple negation rules are applied. That is, a word and its prefix "no", "not", or "without" are combined to one word with the prefix "no_". For example, "without spits" is changed to "no_spits". Third, words without any alphabet are removed. Fourth, the 21 most frequent words are removed because they are function words with little meaning and would othersise take the top positions of learned topics. Lehman et al.'s work [14] used only Unified Medical Language System codes. This enhances the interpretability of the result, but loses valuable information such as patients' social aspects.

5.2 Classifiers

We use cost-sensitive SVMs to handle the imbalance between positive and negative instances (Table 2(c)). Weka[3]'s CostSensitiveClassifier along with SMO configured to return a probability distribution is used in this work. Because of the time-sensitive nature of treatments in an ICU, missing patients in a critical condition may compromise their chances of recovery. In cost-sensitive SVMs, false negatives can be assigned a larger cost than false positives, so that the classifiers are trained not to over-predict negative instances. An alternative to cost-sensitive classifiers could be to resample

[1] http://cs.cmu.edu/~yohanj/research/CIKM15

[2] http://www.physionet.org

[3] http://www.cs.waikato.ac.nz/ml/weka/

negative instances in order to balance them with positive instances. However, this technique suffers from huge variance in effectiveness depending on the resampling result, and we would also miss the opportunity to take advantage of all instances. In contrast, cost-sensitive SVMs can make use of the entire data while providing stable performance scores.

We use 10-fold cross-validation for evaluation. Each fold consists of a training set (81%), a validation set (9%), and a test set (10%). The cost of false positives is fixed to 1, and the training phase explores the space of false negative costs in two rounds to find the best cost using the validation set. The first round explores from 10 to 150 with an interval of 10, finding the cost c with the highest score. The second round searches from $c-9$ to $c+9$ with an interval of 1. The cost that performs the best across the two rounds is selected for the final test. This method has several advantages over searching the entire cost space with an interval of 1. Since performance scores make a unimodal shape in terms of costs, it finds the optimal point faster, and it can circumvent an outlier peak, which may lead to overfitting.

Our evaluation makes use of the Mann-Whitney U test score for comparison, also known as Area Under ROC Curve (AUC). AUC measures how well a trained classifier discriminates positive instances and negative instances and is widely used in this domain.

5.3 Features

Four types of textual features are explored in this work: n-grams, standard topics (i.e., not time sensitive), state-aware topics, and state transitions. These features are associated with different levels of abstractions; n-grams are directly observed, whereas topics and states are inferred.

n-grams: For each patient, unigrams and bigrams are extracted from nursing notes. We empirically decided to use 200 and 100 top n-grams for the mortality group and the survival group, respectively, in terms of pointwise mutual information (PMI). The PMI was calculated on the training data. The decision to use different sizes of n-grams for the two groups comes from the fact that the mortality group, which is much smaller than the survival group, requires more n-grams for high recall. The selected n-grams are merged into the final vocabulary. This process is performed for each prediction term. For each patient, a vocabulary-sized feature vector is made such that each element is set to 1 if the corresponding n-gram appears in the patient's nursing notes written until the prediction time.

Standard topics: Topic distributions learned by LDA are used, using the classification method suggested by Ghassemi et al. [8]. For each patient, topic distributions are extracted from individual nursing notes (not merged notes) written until the prediction time. The extracted topic distributions are averaged and aggregated into one feature vector, whose dimension is equal to the number of topics.

State-aware topics: Document-wise topic distributions ($\theta^m_{t,j}$) learned by STTM are used. STTM estimates one state per time point, thus generating one topic distribution for each time point. The topic distributions of all documents written until the prediction time are aggregated into one feature vector as for standard topics.

State transitions: For each patient, a sequence of states is estimated by STTM. The relative frequencies of states and those of state transitions (pairs of states) are represented as a feature vector. Therefore, if there are S states, the dimension of the feature vector is $S + S^2$. For making this feature, only the latest four time points (i.e., two days) are considered; otherwise, longer sequences tend to include more states and have lower element values. The number four has been chosen empirically.

To simulate the real world prediction task, where topics and states should be estimated as nursing notes are obtained, we estimate the LDA model only on the training set, and the topic distributions of the validation and test data are inferred from the learned topics, using Gibbs sampling. Similarly, we run STTM only on the training set, and a Viterbi algorithm is used to infer an unseen document's state based on the state transition distribution and topic distribution of each state. Once states are finalized, the document's topic distribution is estimated using maximum likelihood estimation.

5.4 Model Parameters

For both LDA and STTM, symmetric Dirichlet priors are used such that α is set to 0.1 and β to 0.001, which fosters sparse distributions over topics and words. γ is set to 1 so that any probability distribution is equally probable for π_c. Gibbs sampling is run for 2000 iterations. The numbers of topics and states are different for each task.

6. EXPERIMENTS

In this section, we first demonstrate the information extracted from the nursing notes by STTM (Task 1). Then, we show improvement in mortality prediction when the extracted temporal information is used (Task 2). Lastly, we evaluate individual features in order to interpret the result of Task 2 and gain insights into the strengths and weaknesses of the individual features (Task 3).

6.1 TASK 1: TEMPORAL INFORMATION EXTRACTION BY STTM

This section is to illustrate and interpret state-aware topics and state transitions learned by STTM over nursing notes on one fold, which will be used for mortality prediction in the tasks evaluated later in the paper. We also demonstrate the kinds of insights STTM is capable of offering. The number of topics and states are set to 10 for the sake of interpretability. Note that interpretability is critical in the medical domain [12], and too many topics and states may make the model hard to interpret. Table 3 and Figure 4 show learned topics and state transitions, respectively.

Table 3 shows the learned topics with manually assigned labels, topical words, and example snippets from nursing notes. Topical words are the words with the highest mutual information with each of the learned topics. These state-aware topics show some differences from standard topics learned by LDA over the same data with the same parameter settings, i.e., on merged nursing notes with 10 topics (not listed here for space). First, there are cases where STTM dedicates two topics that correspond to one standard topic as follows. The admission topics for general patients (T3) and infants (T0) capture the fact that general patients and infants usually follow different state sequences (as seen in Figure 4). Newborn jaundice (T4) and sepsis (T8) can also be explained by different state paths on which they appear. Family visiting (T2) and social activities (T5) share many overlapping words, but family visiting is one of the most critical indicators of patient mortality when it involves discussion of a patient's death. On the other hand, general social activities appear in nursing

Table 3: State-Aware Topics Learned From Nursing Notes

T0. Infant admission	**T5. Social activities**
infant, nicu, delivery, cbc, normal, sepsis, maternal, born, admission, risk, baby, gbs, PHI_DATE, apgars, pregnancy	he, she, that, her, his, have, been, but, had, very, also, will, PHI_LASTNAME, if, need, would, after, about, did, they, it
Admit/transfer note Infant admitted from l&D for sepsis eval. [**Name8 (MD) 63**] MD/NNP note for hx and physical.	NPN Pt left AMA, he felt that he needed to get home to help take care of his daughters
T1. Stability	**T6. Summary**
feeds, cares, infant, well, voiding, continue, parents, feeding, active, stable, support, cc, ra, far, crib, swaddled, mom, fen	am, social, place, blood, progress, due, care, up, time, pm, fluid, npn, per, noted, cxr, id, area, cc, pain, urine, site, currently
continue to encourage po feeds. 4: G/D temps stable in an open crib. alert and active with cares. sleeps well inbetween.	NURSING PROGRESS NOTE 11 PM - 7 AM NO COMPLAINTS HR 80'S...SBP 90-100'S..
T2. Family visiting	**T7. Respiratory support**
family, patient, but, today, pts, sats, team, placed, iv, mg, bp, mask, off, afternoon, night, started, status, able, day, continues	thick, vent, secretions, peep, coarse, sputum, abg, suctioned, yellow, goal, tube, trach, remains, gtt, eyes, amts, skin
npn Pt expired at 7:40pm. Family at bedside with pt. Emotional support given to family at this very difficult time.	Resp Care: Pt continues trached and on ventilatory support with simv 600x14/+5 peep/5 psv/fio2 .4 with acceptable abg
T3. Admission	**T8. Sepsis**
ccu, micu, PHI_DATE, cath, s/p, c/o, PHI_HOSPITAL, admitted, admission, sob, floor, arrival, hct, arrived, ns, wife	npo, settings, fio2, pn, infusing, secretions, cbg, rate, vent, amp, coarse, respiratory, gas, lytes, received, ett, white, simv, cloudy
Pt transfered to [**Hospital1 22**] for ERCP which was done on arrival to MICU.	Respiratory Care Note Pt.continues on SIMV 22/5 R 34 and FIO2 33-44%. BS are coarse. Pt. sx'd for mod.
T4. Newborn jaundice	**T9. Laboratory tests**
bili, infant, cpap, baby, isolette, caffeine, servo, nested, phototherapy, enteral, mom, feeds, cc/kg/day, cc/k/d, will	pain, sbp, gtt, neuro, clear, cough, intact, oob, mg, pulses, foley, neo, bs, effect, ct, bases, given, activity, mae, lopressor, po
#6-O: under single phototherapy, bili pending this am. sl jaundiced. #4-O; temp stable in servo isolette, active and alert	7p-7a Neuro: a+ox3. Pain controlled with Percocet. CV: sbp initially 140's, 2.5mg iv with effect. SR with long PR, rate 80's.

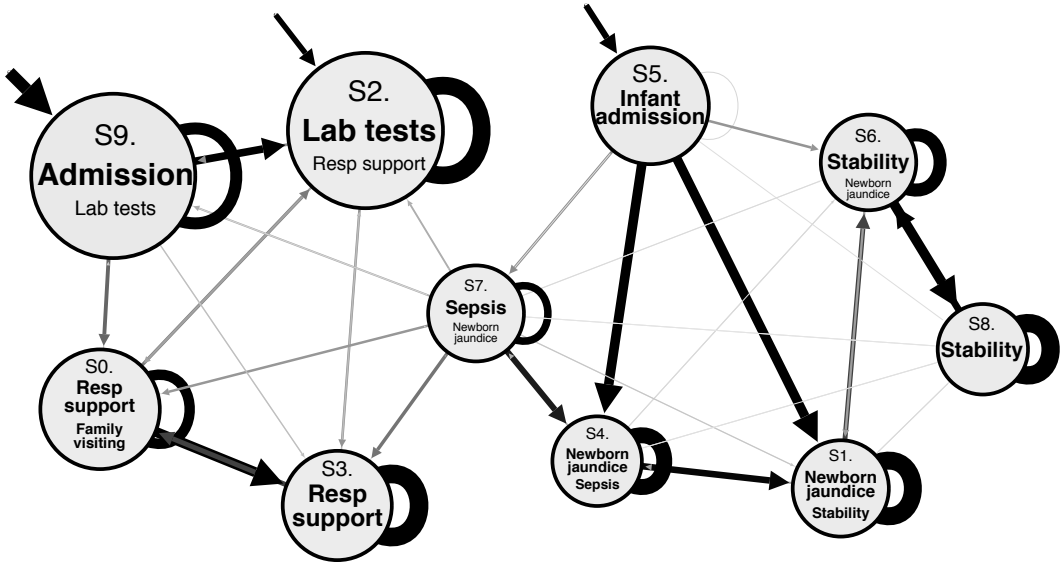


Figure 4: State Transitions Learned From Nursing Notes

notes at anytime. These examples show that STTM splits topics if they appear in different stages of patients. There are opposite cases, too. Stability topic (T1) corresponds to two similar standard topics. Drainage and sounds in organs are topics that appear only in standard topics. These topics may distinguish different symptoms better than state-aware topics do. However, many of these symptoms appear simultaneously in one stage of a patient's journey, and STTM tends to combine them into one topic. For instance, there may be many different outcomes of a patient's stability, but STTM counts all of them as one topic that appears when a patient's condition becomes stable, and dedicates other topics to explain different stages.

Figure 4 is the state transition graph constructed by the learned transition probabilities. Nodes are states and arrows are transitions. Since a state is represented by a topic distribution, node labels reflect the relative proportions of topics. Yet, the labels should not be overemphasized as if each state

covers only those topics in the label. Node sizes reflect the number of patients that are assigned to those states at least once. The thickness of an arrow reflects the transition probability. The arrows coming into S9, S2, and S5 without a source state indicate the probabilities of these states being initial states. Transitions with too small probabilities are not shown in the graph for clarity.

The most notable characteristic is the two big components bridged by S7. The states in the left component are more descriptive of general patients, whereas the states in the right have a large proportion of infant issues. Stability prevails in the right component, which is consistent with the data, wherein neonates are likely to end up being stable in ICUs. The states S9, S2, and S5 are found to be initial states in most sequences. This is reflected by their constituting topics as well, that is, large proportions of general admission (T3) and infant admission (T0). State 2 represents various laboratory tests, which are also common in patients' initial stages. It is interesting that only the infant admission state (S5) is in the right component, indicating the different patterns of NICU patients from other patients. Data analysis confirms that State 5 and thus its connected states capture neonatal patients well; 1725 out of 1727 patients who start from State 5 are neonatal patients, and this accounts for 92% of the entire neonatal patients in the training set.

In summary, STTM captures topics with unique characteristics by taking into account time and state transition. Also, it may not be as straightforward to use other time-aware topic models described in Section 2 to draw a diagram like Figure 4 that provides the whole picture of patient states and state transitions.

6.2 TASK 2: MORTALITY PREDICTION WITH TEMPORAL INFORMATION

This section examines mortality prediction performance meant to validate the insights offered in Task 1. We examine whether temporal information conveyed by state-aware topics and state transitions add any predictive power over standard topics in mortality prediction. The following four combinations of features are compared.

GT: Baseline 50 standard topics. This corresponds to the state-of-the-art method used by Ghassemi et al. [8].

GT+ST: 50 standard topics plus state transitions trained with 10 topics and 10 states.

GT+SA: 50 standard topics plus state transitions and state-aware topics trained with 10 topics and 10 states.

GT+SA+: This feature is the same as GT+SA except that for time points 11-20, it uses state transitions trained on longer sequences. Since most sequences are very short (the median is 2.03 days), STTM is hindered from obtaining good prediction performance for patients with long ICU stays. We therefore train STTM on a subset of training data that includes only those sequences longer than five time points. For prediction at time points 1-10, the original state transitions are used, and at time points 11-20, the newly trained state transitions are used. Therefore, GT+SA and GT+SA+ make the same performance until time point 10.

Table 4 shows the AUC scores averaged across time points. Best scores are in bold, and scores higher than the baseline (GT) are marked with an asterisk if it is statistically significant. Paired t-tests were used for significance tests, and p-values < 0.05 are considered significant as in common mortality prediction tasks [16]. For short-term predic-

Table 4: Performance of Mortality Prediction with Combined Features

	GT	*GT+ST*	*GT+SA*	*GT+SA+*
1-Day	**0.7325**	0.7218	0.7224	0.7235
1-Week	0.7781	0.7772	**0.7811**	0.7784
1-Month	0.7820	*0.7860	*0.7866	***0.7871**
6-Month	0.7882	0.7884	***0.7921**	*0.7912
1-Year	0.7905	0.7912	0.7930	***0.7939**

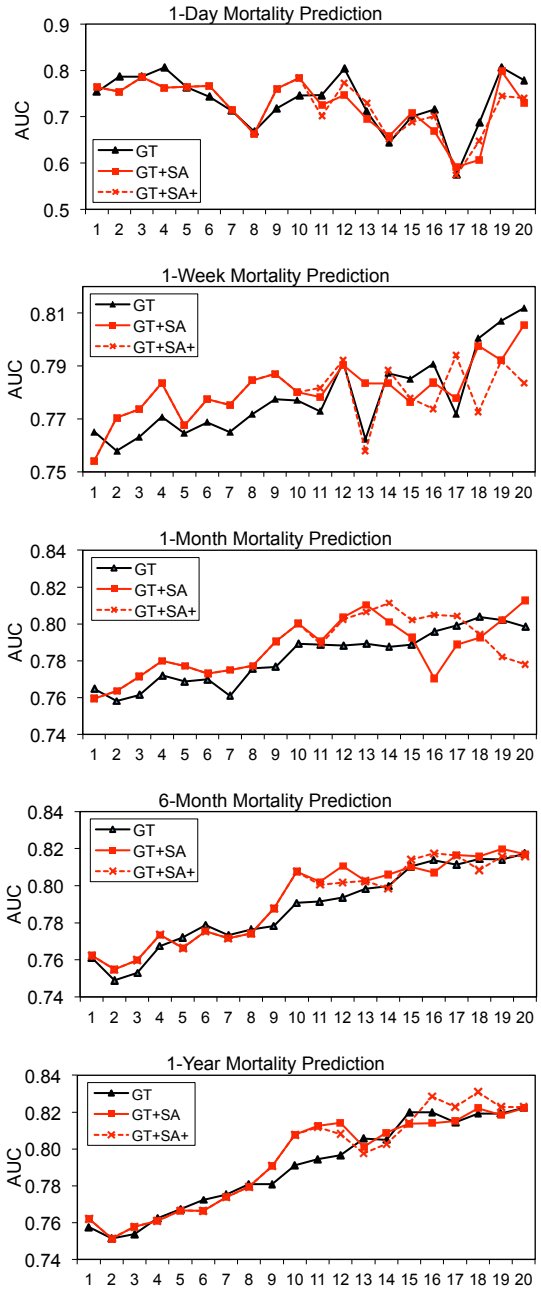


Figure 5: Performance of Mortality Prediction by Prediction Terms and Time Points

tions (1-day and 1-week), incorporating temporal information does not significantly enhance performance. However, for long-term predictions (1-month, 6-months, and 1-year), the temporal information achieves a statistically significant improvement. Training STTM on long sequences turns out to be effective for performance improvement. Figure 5 shows the performance by terms and time points.

Short Term (1-day & 1-week): For 1-day prediction, incorporating temporal information very slightly harms performance, although the effect is not statistically significant. We found no difference in the diversity of states or the frequency of state change between patients who died within one day and those surviving longer. Hence, this result is rooted in the difficulty of predicting patient outcome in the very near future. For instance, it is very hard to tell whether a patient would die today or tomorrow. For 1-week prediction, temporal information starts to help achieve higher performance, although it is not statistically significant yet. There exists a big gap in average scores between 1-day and 1-week. This reflects the difficulty of predicting very short-term consequences. Intuitively, for a patient who is fighting against death right now, clinical measures could be more helpful. This supports the result of previous research that included laboratory and radiology reports [8], which shows higher prediction accuracies for shorter terms.

Long Term (1-month, 6-month & 1-year): For long-term predictions, GT+SA and GT+SA+ outperform the baseline with statistical significance. Training on long sequences is helpful to achieve higher accuracies for 1-month and 1-year predictions. As expected, it enhances prediction performance for later time points, because long sequences help STTM to be trained well for patients who stay in ICUs for a long time. Interestingly, all the combined models (GT+ST, GT+SA, GT+SA+) outperform the baseline for 1-month prediction, but their statistical significance decreases for longer terms. This indicates that temporal information during an ICU stay loses its prediction power as the target future becomes further. This task proves the effectiveness of the temporal information extracted from nursing notes in mortality prediction tasks.

6.3 TASK 3: MORTALITY PREDICTION BY INDIVIDUAL FEATURES

This section measures mortality prediction performance based on each of the four types of textual features: n-grams, standard topics, state-aware topics, and state transitions. The purpose of this task is to take a closer look at what information each feature type captures in mortality prediction. These word-level, topic-level, and state-level features can be seen as being associated with different levels of abstraction. To the best of our knowledge, there is no article that provides a comprehensive view of different types of textual features in mortality prediction tasks. The number of states is fixed to 10 in our experiments.

Table 5 shows the mortality prediction performance of the four individual feature types. Best scores are in bold, and an asterisk is marked if a score is higher than the baseline (StandTopic) or a baseline score is higher than any other scores with statistical significance. N-grams perform poorly compared to topics and state transitions. For 10 topics, state-aware topics perform better than standard topics in most cases, and even state transitions alone are competitive with standard topics. This shows that the learned temporal information is correlated with patient mortality. For 50 topics, the performance of standard topics improves a lot because increasing the number of topics enables to capture more various topics related to mortality. On the other hand, the performance of state-aware topics and state transitions remains almost the same. This is probably because the number of possible topic distributions is restricted to the number of states, preventing documents from having a variety of topic distributions. Although state-related features do not beat standard topics on their own, recall that they enhance performance when used together as shown in Task 2. The performance of state-related features may improve if the number of states is increased at the expense of interpretability. Replacing state parameters θ's with Dirichlet priors α's may also give more freedom to topic distributions.

The rest of this section discusses the performance of individual features. The high markers of mortality and error cases are examined to identify their limitations and possible directions for improvement. For illustration, we pick the classifier at time point 4 (e.g., two days after admission), which is near median of all ICU stays (=2.03 days).

n-grams: To see indicative n-grams, enrichment [8] has been calculated for each n-gram w as $\frac{\sum_n q_{nw} * \mathbf{1}(n)}{\sum_n q_{nw}}$, where n is a patient index, q_{nw} is the feature value, and $\mathbf{1}(n)$ is equal to 1 if patient n died. The enrichment of an n-gram represents the relative frequency of positive instances who have that n-gram, so higher enrichment indicates higher likelihood of mortality. Table 6(a) shows n-grams with the highest and lowest enrichment on average across different terms. Only those n-grams that occur in at least 1% of the entire nursing notes are shown to filter out overfitted n-grams. High mortality markers include descriptions about patient conditions (e.g., "neuro unresponsive" and "mottled"), nurses' actions (e.g., "cmo (comfort measures only)" and "dnr (do not resuscitate)"), medications (e.g., "levophed" and "vasopressin"), and family responses (e.g., "priest" and "meeting held"). High survival markers include a lot of patient conditions such as "no_spits", "active alert", and "ad lib".

Table 5: Performance of Mortality Prediction with Individual Features

		10 Topics				*50 Topics*			
	n-grams	*StandTopic*	*StateTopic*	*StateTrans*	*StateAll*	*StandTopic*	*StateTopic*	*StateTrans*	*StateAll*
1-Day	0.563	0.709	**0.711**	0.709	0.708	***0.733**	0.705	0.703	0.691
1-Week	0.651	0.746	**0.749**	0.746	0.747	***0.778**	0.749	0.737	0.751
1-Month	0.691	0.753	0.759	0.758	***0.762**	***0.782**	0.758	0.749	0.763
6-Month	0.726	0.766	0.767	0.766	***0.771**	***0.788**	0.769	0.757	0.769
1-Year	0.732	0.772	0.771	0.769	**0.773**	***0.790**	0.776	0.760	0.775

Table 6: n-gram Features
(a) High markers of mortality

cmo, neuro unresponsive, poor prognosis, prognosis, priest, fixed, dnr, comfort measures, wishes, dnr/dni, brain, levophed

(b) n-grams with high error contribution

afib, afebrile, nl, will continue, labs, tlc, distended bs, cardiac, abd, code status, right, rate, meds, distended, only,

Table 7: Standard Topic Features
(a) High markers of mortality

T16. Unresponsiveness	neuro, pupils, mm, eyes, left, right, head, drain, sbp, icp
T10. Heart failure medicine	gtt, dr, PHI_LASTNAME, fluid, levophed, bolus, levo, aware, map, started
T29. Pain reliever	propofol, sedated, sedation, fentanyl, versed, peep, gtt, cc/hr, vent, secretions

(b) Topics with high error contribution

T41. Infant delivery	nicu, delivery, infant, cbc, maternal, sepsis, born, admission, gbs, risk
T28. Respiratory infection	settings, fio2, vent, gas, rate, secretion, cbg, map, npo, ett, lytes
T17. TSICU	per, t/sicu, skin, id, tsicu, systems, review, endo, prophylaxis, fx

To examine the reason for the poor performance of n-grams, each n-gram's error contribution has been calculated as $f_w(N_w^{\mathrm{P}} - N_w^{\mathrm{N}})$, where f_w is the feature weight of n-gram w learned by SVM, and N_w^{P} and N_w^{N} are the fractions of positive instances and negative instances who have this n-gram, respectively. Since features that contribute to mortality (survival) have negative (positive) feature weights, the lower the error contribution the better. Table 6(b) shows the n-grams with the highest error contribution. The list includes apparently neutral terms such as "will continue", "labs", "right", and "rate", whose polarity depends on the context. Most of these terms are also very frequent, which significantly harms the prediction performance of n-grams.

Standard Topics (50 topics): We calculated the enrichment of each topic across different terms (Table 7). In this case, the feature values are real numbers between 0 and 1. Unresponsiveness (T16) and taking heart failure medicines (T10) and pain relievers (T29) turn out to be significant signs across all prediction terms. These topics and the n-gram high markers capture similar themes. However, the reduced dimension and the lower sparsity of standard topics significantly improves performance.

The error contribution of the topics shows that infant delivery (T41), respiratory infection (T28), and TSICU (trauma surgical ICU) (T17) contribute to classification error the most. This is reasonable because these topics are not fine-grained enough to the extent that their existence directly leads to mortality. This is the limitation of unsupervised topic models, and splitting these topics may improve performance significantly.

State-Aware Topics (10 topics): Figure 6(a) shows the enrichment of each state-aware topic. Respiratory support (T7), family visiting (T2), summary (T6), and admission (T3) are always in top for all terms. Laboratory tests (T9) are slightly lower. Stability (T1), newborn jaundice (T4), infant admission (T0), and sepsis (T8) have very low enrichment. These topics are reliably indicative of survival.

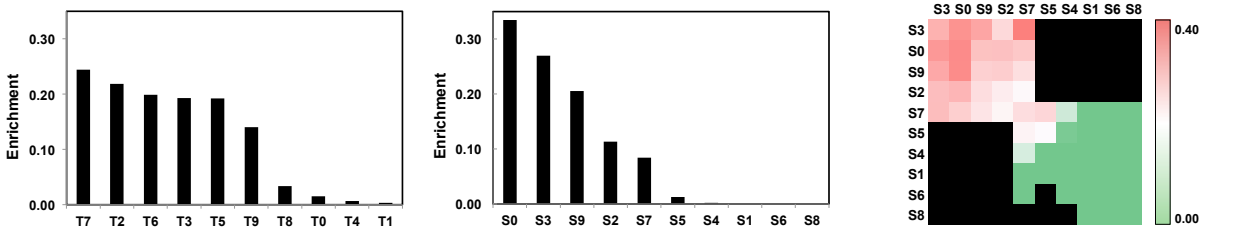
State Transitions (10 topics): Figure 6(b) shows the enrichment of each state. Interestingly, states have higher variance in enrichment than state-aware topics, indicating that states do reflect different mortality possibilities. The state about respiratory support and family visiting (S0) has the highest enrichment, followed by the state about respiratory support without family visiting (S3).

Figure 6(c) shows the average enrichment of state transitions over all terms. Rows are source states and columns are target states. The first five are the states in the left component in Figure 4, and the last five are in the right component. Red and green indicate high and low enrichment, respectively, and black represents no transition found. Patients who move from respiratory problems (S3) to sepsis (S7) and who are admitted (S9) and moved to respiratory support (S0) show the highest enrichment. In contrast, patients who move from respiratory problems (S3) to general lab tests (S2) show relatively low enrichment. State transitions centered around S7 (sepsis) show moderate enrichment. This is consistent with the recent report that mortality rates can be significantly reduced when sepsis is well controlled [1].

Note that high markers differ across time points. For example, high marker topics after five days from admission comprise a lot of neonate-related topics, because most adult patients have been discharged by this time (see Table 2(c)).

7. CONCLUSION

This paper proposes and evaluates a novel approach to mortality prediction using latent temporal information in nursing notes. Our joint model of an HMM and LDA reveals the temporal dynamics of patients' underlying states latent



(a) Enrichment of state-aware topics (b) Enrichment of states (c) Enrichment of state transitions

Figure 6: State-Related Features

within nursing notes. The information—state-aware topics and state transitions—is then leveraged to improve mortality prediction performance. Task 1 shows that the model finds a meaningful trend of patients' state transitions and topics from nursing notes. Task 2 shows that the learned temporal information is beneficial for long term mortality prediction, but not much in short-term prediction. Task 3 suggests that the learned states indeed have different levels of enrichment indicating that the states are related with mortality and survival. In addition, four types of text features are examined both quantitatively and qualitatively, providing a comprehensive view of the roles and limitations of the textual features in mortality prediction tasks.

There are limitations as well. The current version of STTM shows no improvement in mortality prediction when the number of topics is increased from 10 to 50. This may be partly because the number of possible topic distributions is restricted to the number of states. However, we need to try STTM on other data and evaluate in different aspects as well in order to better understand the scalability of STTM. Another limitation is that our approach has no improvement when applied to NICUs and the others separately. This might be due to the reduced data size and the sparsity of the feature space (e.g., 100 possible state transitions). For example, the number of NICU patients is only a fourth of the entire patient population, and the other ICU patients have very short stays (Table 2(c)). More sophisticated approaches are desirable to use temporal information for different ICU types.

8. ACKNOWLEDGMENTS

This research was supported by the Naval Research Laboratory under grant number N00173-09-F-0237.

9. REFERENCES

[1] J. R. Beardsley, C. M. Jones, J. Chou, M. Currie-Coyoy, T. Jackson, and A. Orsborn. Code Sepsis: Improving Sepsis Care; Saving Patients' Lives. *ASHP Best Practices Award*, 2014.

[2] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *ICML '06*, pages 113–120, 2006.

[3] D. M. Blei and P. J. Moreno. Topic segmentation with an aspect hidden Markov model. In *ACM SIGIR '01*, pages 343–348, 2001.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[5] C.-c. Chia, Z. Syed, and A. Arbor. Scalable Noise Mining in Long-Term Electrocardiographic Time-Series to Predict Death Following Heart Attacks. *KDD '14*, pages 125–134, 2014.

[6] M. Cvach. Monitor alarm fatigue: an integrative review. *Biomedical instrumentation & technology*, 46(4):268–77, 2012.

[7] Z. J. Gao, Y. Song, S. Liu, H. Wang, H. Wei, Y. Chen, and W. Cui. Tracking and Connecting Topics via Incremental Hierarchical Dirichlet Processes. In *IEEE ICDM '11*, pages 1056–1061, 2011.

[8] M. Ghassemi, T. Naumann, F. Doshi-Velez, N. Brimmer, R. Joshi, A. Rumshisky, and P. Szolovits. Unfolding Physiological State : Mortality Modelling in Intensive Care Units. *KDD '14*, pages 75–84, 2014.

[9] A. Gruber, M. Rosen-Zvi, and Y. Weiss. Hidden topic Markov models. *AISTATS '07*, pages 163–170, 2007.

[10] L. Hong, B. Dom, S. Gurumurthy, and K. Tsioutsiouliklis. A time-dependent topic model for multiple text streams. In *KDD '11*, page 832, 2011.

[11] L. Hong, D. Yin, J. Guo, and B. D. Davison. Tracking trends: incorporating term volume into temporal topic models. In *KDD '11*, page 484, 2011.

[12] A. Kalogeratos, V. Chasanis, G. Rakocevic, A. Likas, Z. Babovic, and M. Novakovic. Mining Clinical Data. In G. Rakocevic, T. Djukic, N. Filipovic, and V. Milutinović, editors, *Computational Medicine in Data Mining and Modeling SE - 1*, pages 1–34. 2013.

[13] L.-W. Lehman, R. Adams, L. Mayaud, G. Moody, A. Malhotra, R. Mark, and S. Nemati. A physiological time series dynamics-based approach to patient monitoring and outcome prediction. *IEEE journal of biomedical and health informatics*, PP(99):1, 2014.

[14] L.-w. Lehman, M. Saeed, W. Long, J. Lee, and R. Mark. Risk stratification of ICU patients using topic models inferred from unstructured progress notes. *AMIA Annual Symposium proceedings*, 2012:505–11, 2012.

[15] Q. Mei and C. Zhai. Discovering evolutionary theme patterns from text. In *KDD '05*, page 198, 2005.

[16] R. Menéndez, R. Martínez, S. Reyes, J. Mensa, X. Filella, M. A. Marcos, A. Martínez, C. Esquinas, P. Ramirez, and A. Torres. Biomarkers improve mortality prediction by prognostic scales in community -acquired pneumonia. *Thorax*, 64(7):587–91, 2009.

[17] M. J. Paul. Mixed Membership Markov Models for Unsupervised Conversation Modeling. In *EMNLP-CoNLL '12*, pages 94–104, 2012.

[18] Z. Syed, B. M. Scirica, S. Mohanavelu, P. Sung, E. L. Michelson, C. P. Cannon, P. H. Stone, C. M. Stultz, and J. V. Guttag. Relation of death within 90 days of non-ST-elevation acute coronary syndromes to variability in electrocardiographic morphology. *The American journal of cardiology*, 103(3):307–11, 2009.

[19] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[20] H. Wang, D. Zhang, and C. Zhai. Structural Topic Model for Latent Topical Structure Analysis. *ACL '11*, pages 1526–1535, 2011.

[21] X. Wang and A. McCallum. Topics over time: a non-Markov continuous-time model of topical trends. In *KDD '06*, page 424, 2006.

[22] Y. Wang, E. Agichtein, and M. Benzi. Tm-lda: Efficient online modeling of latent topic transitions in social media. In *KDD '12*, pages 123–131, 2012.

[23] J. Wiens, E. Horvitz, and J. V. Guttag. Patient Risk Stratification for Hospital-Associated C. diff as a Time-Series Classification Task. In *Advances in Neural Information Processing Systems*, pages 467–475, 2012.

[24] J. Yang, J. McAuley, J. Leskovec, P. LePendu, and N. Shah. Finding progression stages in time-evolving event sequences. In *WWW '14*, pages 783–794, 2014.

[25] J. Zhang, Y. Song, C. Zhang, and S. Liu. Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora. In *KDD '10*, page 1079, 2010.

[26] J. Zhu and E. P. Xing. Conditional topic random fields. *ICML '10*, pages 1239–1246, 2010.