Learning to Summarize Electronic Health Records Using Cross-Modality Correspondences

Jen J. Gong JENGONG@MIT.EDU
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA, USA

John V. Guttag GUTTAG@MIT.EDU
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA, USA

Abstract

Electronic Health Records (EHRs) contain an overwhelming amount of information about each patient, making it difficult for clinicians to quickly find the most salient information. Accurate, concise summarization of relevant data can help alleviate this cognitive burden. In practice, clinical narrative notes serve this purpose during the course of care, but they are only intermittently updated and are sometimes missing information.

We address this problem by learning to generate topics that should be in summaries of structured health record data at any point during a stay. We use the detailed, high-dimensional structured data to predict existing clinical note topics. Our model can generate topics based on structured health record data, even when a real note does not exist. We demonstrate that using structured data alone, we are able to generate note topics comparable to the performance of using prior notes alone. Our method is also capable of generating the first note in the stay.

We demonstrate that our predicted topic distributions are meaningful using the downstream task of predicting in-hospital mortality. We show that our generated note topic vectors perform comparably or even outperform topics from the actual notes on predicting in-hospital mortality.

1. Introduction

Electronic Health Records (EHRs) contain an overwhelming amount of information about each patient, making it difficult for clinicians to quickly find the most salient information at various points during an admission. Information overload can also result in health care providers missing important information during the course of care (Singh et al., 2013). Accurate, concise summarization of relevant data can help alleviate this cognitive burden.

Clinical narrative notes serve this purpose during the course of care. They help clinicians summarize and identify the most relevant aspects of the deluge of available data about each patient, and facilitate communication among care teams (Kuhn et al., 2015). However, since clinical notes are written at infrequent intervals, information from the most recent note can quickly become outdated. This is particularly true in critical care settings, where patient state can suddenly change and interventions are frequently administered. Missing information during communication between care team members can lead to adverse events (Arora et al., 2005). Methods for assisting care team members in writing summaries of patient state and the course of care could help address potential errors of omission.

In this paper, we propose a system that generates relevant patient- and time-specific *topics* from structured health record data. We use a supervised modeling approach to learn correspondences between detailed, high-dimensional structured data and existing clinical notes. We model each note as a distribution over topics using latent Dirichlet allocation (LDA) (Blei et al., 2003). These topics have been shown to capture relevant patient subtypes, and are predictive of adverse outcomes such as mortality (Ghassemi et al., 2014) and interventions (Suresh et al., 2017). We then use our model to generate topic-based summaries of structured health record data.

The contributions of this work are:

1. We present a supervised framework to learn correspondences between high-dimensional structured EHR data elements and low-dimensional topic representations of clinical notes over the course of a patient stay. This model can be used to summarize patient care and physiology – even when a note was never written.

2. We evaluate the generated topic distributions. We show that the generated topic distributions reflect changes in patient state earlier than recorded clinical notes, and reflect meaningful correspondences between topics and relevant structured items.

3. We show that using structured data alone to predict the next note performs similarly to using all prior notes when they exist. In addition, structured data can accurately predict the first note in a patient stay, when a model using only the notes data has no information. We show that combining structured data and notes can improve predictions over either one alone when prior notes exist.

4. We evaluate topics generated from the structured data alone by evaluating performance on a downstream prediction task: in-hospital mortality. We demonstrate that a model built using only our predicted notes leads to comparable performance to using a model built from the actual notes.

Technical Significance: To our knowledge, our work is the first that proposes to use high-dimensional structured EHR data to generate topics that may be missing in clinical notes. We propose a novel supervised method, using existing clinical notes as labels to learn meaningful correspondences between summaries written by clinicians and structured health record data.

Clinical Relevance: Clinical notes are used at the point of care to summarize patient state; however, they are intermittently updated, and are sometimes missing information. Our model can be used to generate topics from structured health record data that should be in a patient's clinical note. These topics can be used as a checklist for clinicians while they are writing the note.

We first discuss related work in Section 2. Next, we describe the data in Section 3 and our data processing methods in Section 4. We describe our methods in Section 5, and our experimental results in Section 6. Finally, we summarize our findings in Section 7.

2. Related Work

2.1. Summarizing Health Record Data

A great deal of work has investigated how to summarize structured health record data in a more accessible manner. Some works have used visualization interfaces (Monroe et al., 2013; Plaisant et al., 1996; Hirsch et al., 2014). Others have used natural language to generate descriptions of structured time-series (Goldstein and Shahar, 2016; Portet et al., 2009; Hunter et al., 2008). Pivovarov and Elhadad (2015) contains a comprehensive summary of techniques for summarizing health record data. In contrast to these works, our goal is to automatically learn correspondences between structured data and existing summaries written during the course of care.

2.2. Clinical Note Time-Series

Our work leverages cross-modal data relationships to predict notes at times when they are not usually written. Ghassemi et al. (2015) handles the problem of missing notes by learning a multi-task Gaussian Process (MTGP) over the time-series of clinical notes. The authors do not evaluate the ability of the MTGP to forecast notes. They instead demonstrate the utility of the MTGP parameters for downstream prediction tasks (e.g., in-hospital mortality). In contrast, we are interested in the task of forecasting topic membership of missing clinical notes, to generate summaries of care even when they are not present.

Jo et al. (2015) models evolving patient state from nursing notes using a model that integrates a hidden Markov model (HMM) and latent Dirichlet allocation (LDA). This model captures changing patient dynamics (and therefore changing topic memberships) over time, but does not consider the additional value of structured health record data for generating clinical note topics.

2.3. Integrating Clinical Data Modalities

In this work, we consider physiological time-series, clinical events, and clinical notes. Each modality of data has been shown to be successful in predicting clinical outcomes such as mortality (e.g., Che et al. (2018); Gong et al. (2017); Ghassemi et al. (2014)) and intervention administration (Ghassemi et al., 2017; Wu et al., 2017). In addition, multimodal EHR data have been integrated, primarily for the tasks of 1) patient phenotyping (e.g., Ho et al. (2014); Pivovarov et al. (2015); Henao et al. (2016)), and 2) clinical outcome prediction (e.g., Suresh et al. (2017); Huddar et al. (2016); Caballero Barajas and Akella (2015)). While we demonstrate the utility of our learned correspondences in downstream predictive tasks, we are primarily focused on the task of learning a correspondence between the structured data time-series and a note summarizing patient status and the care process.

3. Data

We use data from MIMIC-III (v 1.4), a publicly available dataset consisting of data collected in the intensive care units (ICUs) at the Beth Israel Deaconess Medical Center over the years 2001 - 2012 (Johnson et al., 2016). MIMIC-III contains data from two EHR systems: 1) CareVue (2001-2008) and 2) MetaVision (2008-2012). Because the encodings of clinical events differed significantly between the two EHR systems in MIMIC-III, we considered only data from the latter version (MetaVision, 2008-2012).

In the following sections, we describe our cohort selection criteria, and each of the three modalities of clinical data we considered: 1) *clinical event sequences*, 2) *physiological time-series*, and 3) *clinical notes*.

3.1. Cohort Selection

We considered patients ≥ 15 years of age. We used each patient's first ICU stay, to avoid multiple admissions from the same patient. Patients who died, were discharged, or had a note of "comfort-measures only" within 12 hours of ICU admission were removed from the study. Patients missing any of the three modalities of data were also removed, reducing our patient population from roughly $15,000$ patients to 6,360 patients. This reduction was primarily a result of patients without regular physician and nursing notes. The differences between patients with and without notes are detailed in Appendix A. Patients with missing notes are not noticeably different from patients with notes in length of stay in the ICU, presence in different care units, or admission status. However, mortality rate was elevated in patients with missing notes. We divided the remaining patients into a 60/20/20 training/validation/test split. These divisions are described in Table 1.

3.2. Data Modalities

Events: Time-stamped clinical events were extracted for each patient. These included procedures, lab tests and results, input/output events (e.g., medications, fluids), microbiology tests, observations noted in the chart, and service changes. Unique (item, text) pairs were considered as distinct events, as in Gong et al. (2017).

Physiological Time-Series: We extracted for each patient 31 vital signs and lab values from the database, as in Ghassemi et al. (2017); Suresh et al. (2017); Wu et al. (2017). These features included diastolic, systolic, and mean blood pressure, heart rate, respiratory rate, temperature, height, weight, white blood cell count, pH, albumin, anion gap, bicarbonate, bilirubin, blood urea nitrogen, chloride, creatinine, fraction inspired oxygen, glucose, hematocrit, hemoglobin, INR, lactate, magnesium, oxygen saturation, partial thromboplastin time, phosphate, platelets, potassium, prothrombin time, and sodium. These signals differ in the frequency and regularity at which they are sampled; e.g., whereas vital signs are sampled regularly in the ICU, lab tests are ordered intermittently.

Clinical Notes: We extracted time-stamped physician, nursing, and general notes for each patient. These note categories summarize care provided during the ICU stay. We excluded other categories of notes, such as radiology reports, echo reports, and ECG reports. Discharge summaries were excluded because they summarize the stay after it is over. Figure 1 shows the number of admissions with notes at each hour of the ICU stay, aligned on midnight of the day of ICU admission.

Table 1: Cohort and training/validation/test data split descriptions.

	Train	**Validation**	**Test**
Number of Patients	3816	1272	1272
Number of Notes	111,938	34,553	38,747
In-Hospital Mortality	7.0%	7.5%	7.2%
Mean (std) LOS in ICU (days)	2.5 (1.9)	2.4 (1.8)	2.5 (2.0)

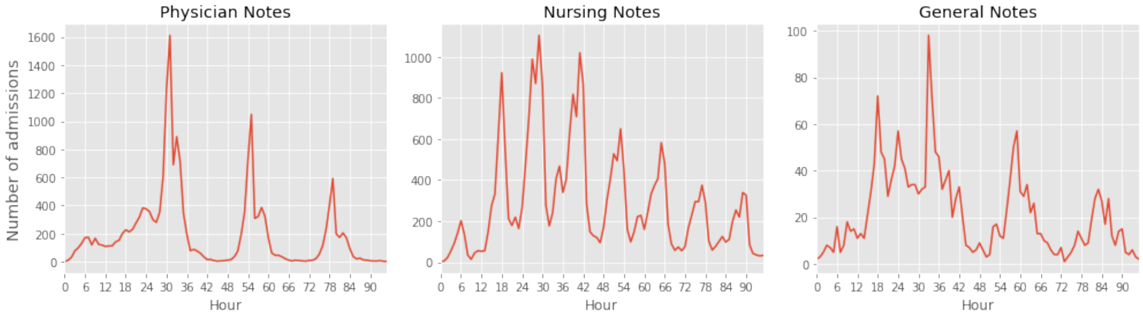

Figure 1: Timing of physician notes peaks at 6 a.m. in the morning. Timing of nursing notes is more irregular than physician notes, but exhibits regular inter-event intervals of approximately 6 hours.

4. Data Processing

All data were aligned to midnight on the day of ICU admission, to preserve time-of-day characteristics, and discretized to the hour. All admissions were padded or truncated to 96 hours from midnight of the first day of ICU admission. The following sections describe processing details for each data modality.

4.1. Events

We discretized the times of events to the hour, from midnight on the day of ICU admission. Events that occurred in the same hour were represented with a binary bag-of-events (BOE) vector, indicating whether or not each event occurred in that hour. We considered two types of events: 1) point events, that occurred at a single point in time, and 2) duration events, which were specified with a start and stop time. Events that spanned a duration of time were 1 between the start and stop times, and 0 otherwise. Point events were 1 if the event was present during the hour and 0 otherwise. The events tensor was then constructed by building this BOE vector over time. Events that occurred in fewer than three unique admissions in the training data were filtered out. In total, we considered 6,556 kinds of events.

4.2. Physiological Time-Series

Continuous-valued vital signs and lab test measurements were binned to the hour by taking the median of the values in each hour. The hourly values were then discretized by taking the z-score, rounding to the nearest integer, and mapping outliers ($|z| > 4$) to -4 and 4, following the procedure used in Suresh et al. (2017) and Wu et al. (2017). The means

Table 2: Top 5 and bottom 5 topics by enrichment for in-hospital mortality.

Topic	**High Enrichment for In-Hospital Mortality**
14	family, care, dnr, support, daughter, dni, son, comfort, morphine, social
37	hypotension, line, shock, sepsis, levophed, cvp, fluid, bp, pressors, map
16	liver, cirrhosis, lactulose, transplant, encephalopathy, ascites, hepatic, varices, sbp, albumin
25	spontaneous, rr, min, set, vt, tube, ventilator, peep, mode, ve
36	intubated, sedation, vent, propofol, abg, extubation, sedated, fentanyl, wean, respiratory
Topic	**Low Enrichment for In-Hospital Mortality**
15	etoh, abuse, ciwa, withdrawal, alcohol, pancreatitis, valium, scale, thiamine, seizures
43	pain, control, chronic, acute, continue, prn, dilaudid, morphine, po, iv
42	valuables, transferred, rate, pmh, weight, heart, bp, total, sent, money
13	present, pulse, min, extremities, mmhg, current, regular, rhythm, insulin, chest
38	cabg, artery, wires, coronary, bypass, temporary, graft, svg, avr, valve

and standard deviations of all of the features were determined across all admissions in the training and validation data. These features were then binarized. An additional bin was added for each variable to indicate a missing value.

4.3. Clinical Notes

We filtered out a set of pre-defined clinical stop words (e.g., patient, report, pt, admission, discharge, etc.), as well as tokens that occurred in fewer than 3 documents or more than 95% of documents. Additionally, punctuation and numerical values were filtered out. We used latent Dirichlet allocation (Blei et al., 2003) to reduce the dimensionality of the clinical notes from a $> 47K$ vocabulary to a distribution over 50 topics. Topic models were trained using gensim (Řehůřek and Sojka, 2010). For each patient, the topic distribution at each hour was computed by taking the average of the topic distributions for all notes in that hour.

Table 2 describes the top five and bottom five topics by enrichment for in-hospital mortality. Enrichment was computed using the training data by taking the average topic probability for each topic across all notes, weighted by the outcome of the patient the note was written about, as in Marlin et al. (2012). The full set of topics is described in Appendix B.

5. Methods

5.1. Learning Correspondences

To learn correspondences between the structured clinical data and the clinical notes, we use a supervised deep learning approach that leverages the temporal nature of the structured data and the clinical notes.

5.1.1. Network Architecture

The *struct2note* model uses all structured data up to and including the hour of the note of interest to predict topics for a clinical note. We compare against two other models that

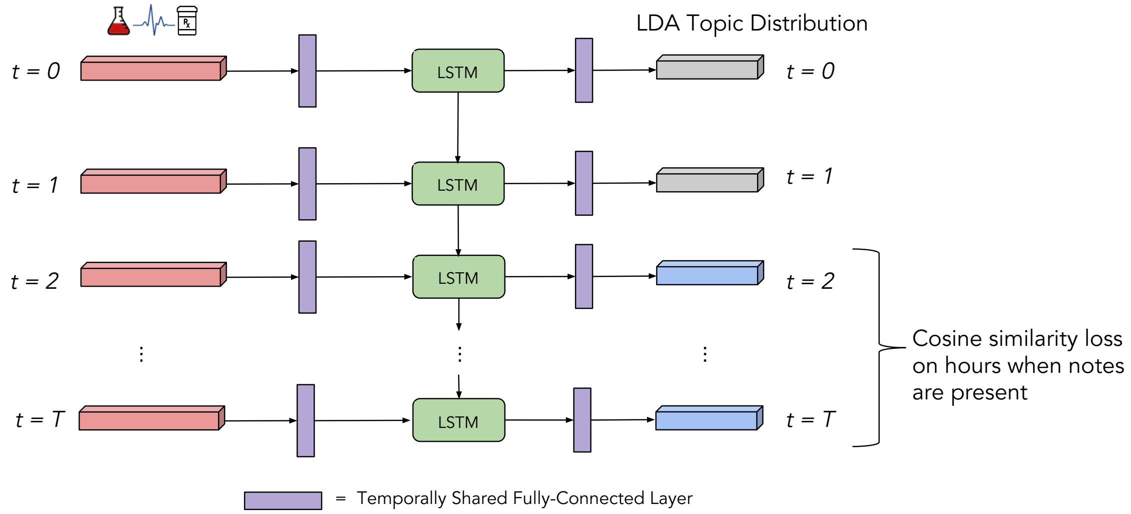
LDA Topic Distribution
t = 0
t = 1
t = 2
t = T
LSTM
Cosine similarity loss on hours when notes are present
= Temporally Shared Fully-Connected Layer

Figure 2: Model architecture for structured data (*struct2note*). The network is shown unrolled over time. Sparse, high-dimensional time-series of structured data are first passed through a fully-connected layer shared over time to get a dense embedding. The time-series are then encoded using an LSTM. The topic distribution for the note at each time step is predicted with a fully-connected layer (shared over time) with a softmax activation. During training, the loss was computed on hours when notes were present.

use prior notes: 1) *notes2note* uses all prior notes to make a prediction, and 2) *struct-notes2note* uses prior notes and structured data. Figure 2 diagrams our model architecture for *struct2note*.

In *struct2note*, a temporally shared fully-connected embedding layer with a rectified linear activation function maps the structured data at each time step from a sparse, high-dimensional feature space to a low-dimensional dense embedding space. This captures relationships between co-occurring events at each time-step. We use a long short-term memory (LSTM) network to capture the temporal patterns in the structured data (Hochreiter and Schmidhuber, 1997). LSTMs have been shown to encode temporal patterns that are effective in predicting interventions and identifying patient diagnoses (Suresh et al., 2017; Lipton et al., 2016). Finally, a temporally shared fully-connected layer with a softmax activation outputs predicted probabilities for the 50 topics at each time step. *notes2note* uses a similar architecture, but because the topics are already a dense embedding space, we do not need an embedding layer. *struct-notes2note* combines both data modalities by concatenating the topic distribution tensor with the embedded structured data tensor as the input through the LSTM.

While we choose to use LSTMs in this work, LSTMs, and neural networks more generally, are not the only method for learning such supervised correspondences. We use these models to demonstrate the feasibility of learning meaningful correspondences between structured health record data and clinical note topics in a supervised framework.

5.1.2. Loss Function and Evaluation Metric

To compare predicted topic distributions with the true topic distributions, we use *cosine similarity*. Cosine similarity is the normalized dot product between two vectors:

$$C(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|}. \tag{1}$$

This measure takes a maximum value of 1 when **u** and **v** are parallel, a value of 0 when **u** and **v** are orthogonal, and a minimum value of -1 when **u** and **v** are anti-parallel. In our application, the minimum value the cosine similarity measure can take is 0, because we are comparing two probability distributions (all elements are non-negative). Cosine similarity is an appropriate loss function because it evaluates how close **u** and **v** are in directionality, rather than in magnitude. Because our topic distributions always sum to 1, magnitude is not important in assessing the differences between the topic distribution of the actual note and the predicted topic distribution. Cosine similarity has been used in prior work to evaluate differences between dense embeddings of words (Mikolov et al., 2013). We use cosine similarity both as the loss function during training, and as an evaluation metric to determine how close our predicted topic distributions are to the true ones.

Clinical notes are not present at every time step. The cosine similarity loss is only considered at time-steps when notes are present. When prior clinical notes are used as input to the *notes2note* and *struct-notes2note* models, notes are forward-filled with the most recent note up until the latest of time of death, discharge, or the final note. Time-steps where input data are not present (e.g., prior to ICU admission on the first day) are masked out.

5.1.3. Training and Implementation

We implemented our models using Keras 2.1.3 with Tensorflow backend (1.5.0) (Chollet et al., 2015). The size of the first temporally shared fully-connected layer for embedding the structured data was set to 30 units, and a grid search from 8 to 256 in multiples of 2 was performed to choose the LSTM hidden layer size. All models were chosen based on the validation loss. The sizes of the LSTM hidden layers for the final models are detailed in Appendix C.

5.2. In-Hospital Mortality Prediction

To demonstrate that our predicted note topics capture meaningful aspects of patient care and state, we predict in-hospital mortality using models trained on 1) existing clinical notes, and 2) the predicted clinical notes. In-hospital mortality is often used as a proxy for patient severity of illness. We use the network architectures from Section 5.1, replacing the topics-over-time output tensor with a binary tensor indicating the outcome for that patient at that time.

We predict whether or not in-hospital mortality occurs at least 24 hours after the hour the prediction is made. We define the outcome using the earliest of the patient's time of death or a note of "comfort-measures only" (CMO). When a patient is declared CMO, few (if any) interventions are made, and the prediction is no longer relevant to the course of care. At each hour, a prediction is made for each patient. Predictions for patients who are discharged or die prior to the hour of prediction or within the 24 gap period are excluded from the loss at that time step. Models trained for in-hospital mortality were further restricted compared to the training, validation, and test sets described in Table 1. Specifically, patients who died, were discharged, or had a note of CMO in the first 24 hours of the ICU stay were filtered out from model training and evaluation.

6. Results

6.1. Predicting the Next Note

To evaluate our model's ability to predict topic vectors for existing clinical notes, we compared against two baselines: 1) *prior note* topic membership, where we used the most recent note topic membership to predict that of the current note, and 2) *average note* topic membership, where we used the average note topic membership from the training data to predict the topic membership of each note in the test data.

Table 3 shows the aggregate prediction results of each model on the notes in the test data. Performance is broken down by notes with prior notes ($n = 9290$), and notes without prior notes ($n = 1272$). We evaluated statistical significance of the difference between the *average* performance of each model across all notes for *each patient*. We evaluated differences in model performance at the patient level, rather than at the note level, because notes belonging to the same patient are not independent. We used a paired t-test with a significance level of 0.001.

Using all prior notes (*notes2note*) and using structured data (*struct2note*) performed comparably well in predicting the next note (mean cosine similarity of 0.63, $p = 0.006$). On the task of predicting the first note in a stay, *notes2note* performed comparably to taking the average note from the training data (cosine similarity of 0.41 vs. 0.42). This makes sense, as the *notes2note* model had no additional information to take advantage of. However, the *struct2note* model was able to predict the first note in the stay with a cosine similarity of 0.61, significantly outperforming the *notes2note* model ($p < 1e - 200$).

In addition, integrating the structured data and prior notes to predict the next note outperformed using either modality on its own (mean cosine similarity of 0.66 vs. 0.63, $p < 1e{-}46$). When no prior note existed (and therefore *struct-notes2note* had no additional

Table 3: Cosine similarity performance of different models on the test set. Mean, standard deviation, and quartiles of performance are shown, broken down by notes where a prior note existed, and notes where no prior note existed.

	Notes with prior notes (9290)				Notes without prior notes (1272)			
		Quartiles				Quartiles		
	Mean (std)	25%	50%	75%	Mean (std)	25%	50%	75%
notes2note	0.63 (0.19)	0.50	0.65	0.78	0.41 (0.09)	0.35	0.42	0.48
struct2note	0.63 (0.21)	0.49	0.66	0.80	**0.61** (0.17)	0.49	0.62	0.74
struct-notes2note	**0.66** (0.21)	0.53	0.69	0.82	**0.61** (0.17)	0.49	0.62	0.73
Prior note	0.39 (0.29)	0.15	0.31	0.62	–	–	–	–
Average note	0.40 (0.09)	0.34	0.41	0.46	0.42 (0.09)	0.36	0.42	0.48

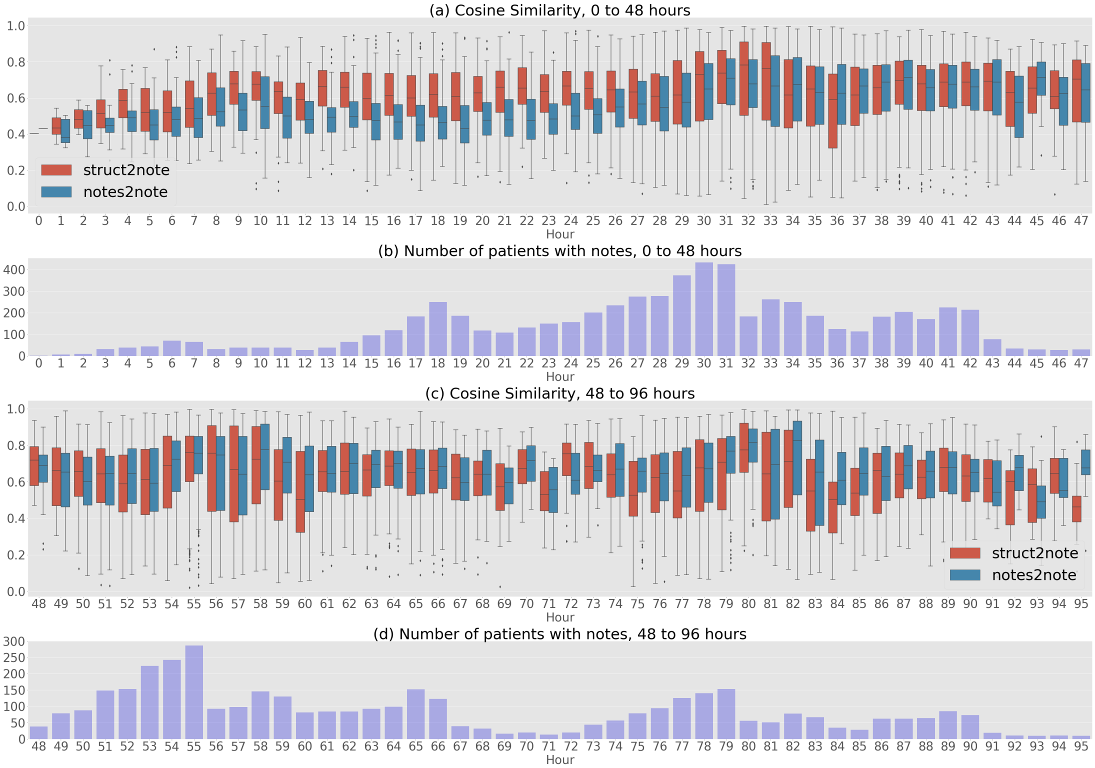

Figure 3: *struct2note* and*notes2note* cosine similarity performance on predicting topics of clinical notes are shown in (a) (0 to 48 hours) and (c) (48 to 96 hours). Number of patients with a note at each hour is shown in (b) (0 to 48 hrs) and (d) (48 to 96 hrs).

information compared to *struct2note*), the average cosine similarities across notes were similar between the two models (0.61).

Because notes have differing availability over time, we investigated the performance of these models on notes at different hours during the ICU stay. The differences in performance between *struct2note* and *notes2note* are shown in Figure 3(a) and (c). The number of patients with a note at each hour is shown in Figure 3(b) and (d).

In the early hours of the ICU stay (0 to 30), the structured data outperforms using the notes. Since there are very few notes available at this time, it is challenging for the *notes2note* model to make meaningful predictions. This performance improvement drops off around hour 30, or 6 a.m. on the second day of the patient's stay in the ICU. Recall from Figure 1 that physician notes are recorded regularly around 6 a.m. each day. At these times, the availability of notes grows, and predictive accuracy of the note prediction models increase. The improvement of using structured data rather than prior notes becomes marginal at later hours of the stay (48-96), when more notes are available.

6.2. Outcome Prediction

We evaluated the note predictions generated from the structured data alone (*struct2note*) by training supervised networks using 1) actual notes and 2) predicted notes for predicting in-hospital mortality.

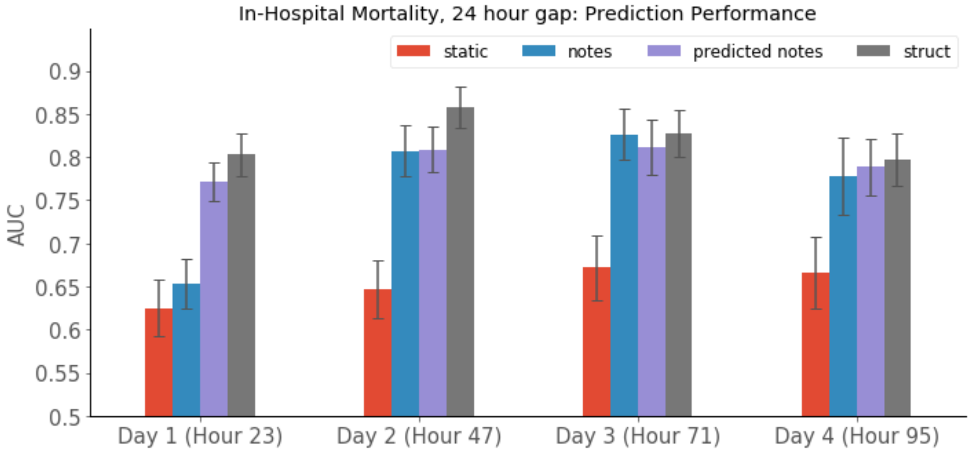

Figure 4: AUC using different data modalities to predict in-hospital mortality in the final hour of each day in the ICU (23, 47, 71, 95 hours). Error bars indicate standard deviations computed across 100 bootstrapped samples.

We evaluated performance in terms of the Area Under the Receiver Operating Characteristic Curve (AUC). We evaluated statistical significance by evaluating model performance on 100 bootstrapped sets for each model. A paired t-test was performed between the bootstrapped AUCs for a pair of models, at a significance level of 0.001. Bootstrapped samples were constructed so that the outcomes were represented in the same incidence as in the original test set. We also trained models using 1) static demographic characteristics such as age, gender, admission type, and first care unit and 2) structured data (events and physiological time-series) as performance baselines. Models utilizing the static data used a fully-connected layer (since the static data do not change over time). All other models used similar model architectures to those described earlier.

The results are shown in Figure 4. We show performance results at the last hour of each day (11 p.m.), when information from the course of the day can be taken into account. Our predicted note topic distributions performed comparably to the actual notes at hours 47 and 95 ($p = 0.78$ at hour 47 and $p = 0.46$ at hour 95). At hour 71, the difference in performance between the predicted note topics (AUC = 0.81) and the actual note topics (AUC = 0.83) was statistically significant ($p < 1e-5$), but not large. In addition, the predicted note topics significantly outperformed the actual ones at hour 23 ($p < 1e-50$).

These performance results indicate that our method of learning correspondences between structured health record data and topic distributions of existing clinical notes allowed us to generate meaningful topics that capture changes in patient state. Importantly, although the predicted note topic distributions do not incorporate *any* of the existing notes, they achieve predictive performance comparable to the topics of the actual notes in downstream prediction tasks.

6.2.1. Visualizing Correspondences

To qualitatively evaluate the learned correspondences,we identified individuals with high presence of certain topics and visualized structured data elements with meaningful relation-

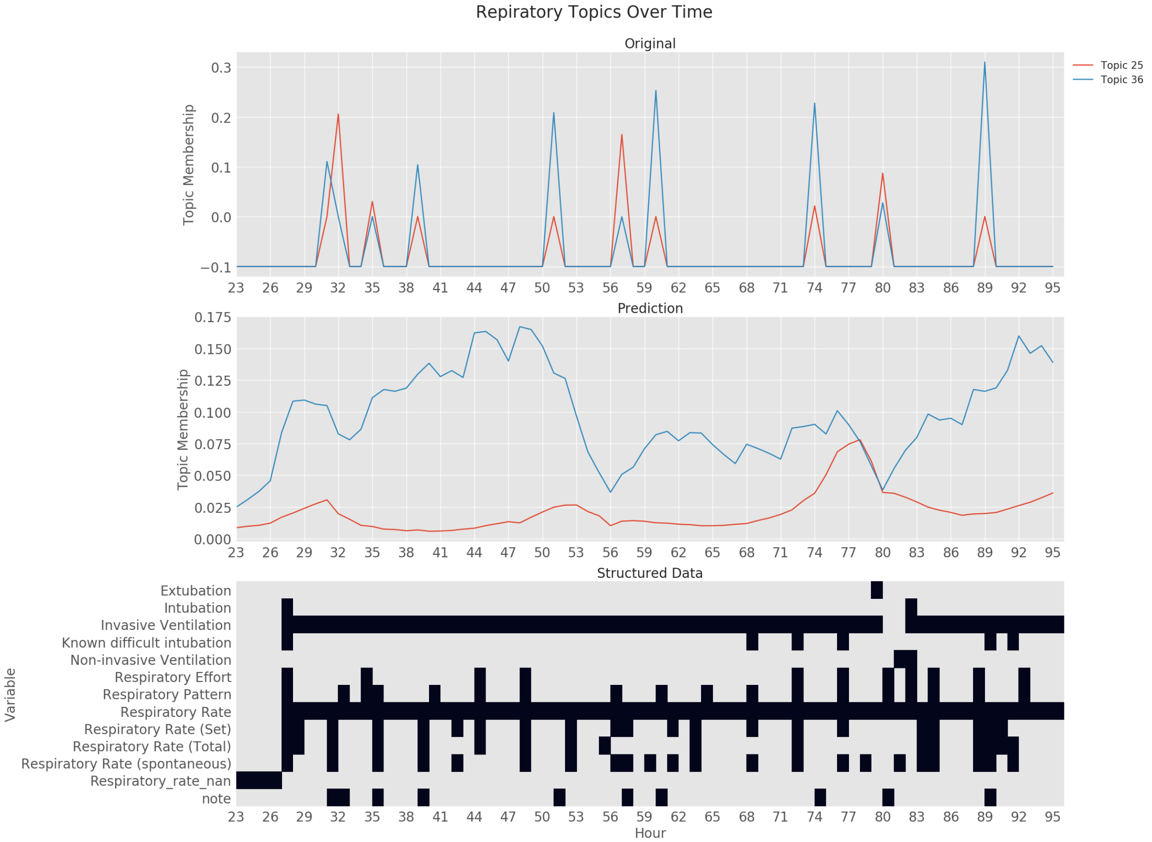

Figure 5: Correspondences between topic distributions of ground truth notes (top), predicted topic distributions (middle), and structured health record data (bottom) for a single admission. Topic membership values are shown as negative when no note was present. Topics corresponding to intubation and respiratory status (25 and 36) are shown, along with structured data elements pertaining to respiratory status and ventilation.

ships to those topics. Figure 5 shows the original topic distributions over time for topics corresponding to intubation or respiratory status (topics 25 and 36). This 88 year-old patient was admitted to the ICU shortly after 11 p.m. (hour 23). Her admission status was "emergency." She died in the hospital, 8 days after admission.

This patient was intubated shortly after ICU admission, at around 4 a.m. (hour 28). Whereas the original note only indicates a rise in corresponding topic membership around hour 31, our predicted note topics show an immediate rise in topic 36. This indicates that our predicted note topics are able to capture changes in patient state before the actual notes are recorded. This occurs again at hour 81, when the patient is extubated and then intubated again shortly after. While the predicted topics show an immediate rise in Topic 36, the note was not written until 8 hours later, at hour 89.

This example demonstrates that our method enables learning meaningful correspondences between the high dimensional structured EHR data and clinical summaries written during the course of care. In addition, we note that while our predicted topics did not always accurately represent the true topic distributions of notes (e.g., at hour 56), they still reflect meaningful correspondences with the structured data. This suggests that even if cosine similarity between the predicted note and the true note is low, our predictions might offer useful suggestions regarding topics that might be missing from the recorded notes.

7. Discussion

In this work, we proposed a method to *learn* to generate meaningful topic summaries from structured patient health record data. We used existing summaries written by clinical care team members to learn correspondences between structured health record data and the topics underlying clinical notes. We demonstrated that using structured data alone, we are able to generate note topics with an average cosine similarity to actual notes of 0.63, comparable to the performance of using prior notes alone. Integrating structured data with prior notes results in an average cosine similarity of 0.66. Using the structured data, we are also able to generate the topics of the first note in the stay with an average cosine similarity of 0.61.

We also demonstrated that our generated topics are able to predict clinical outcomes such as in-hospital mortality with comparable performance to topic distributions of actual notes written by care team members. We additionally presented a qualitative example of correspondences between structured data elements and changes in topic distribution.

Inherent to our approach is an assumption that clinical notes are *good* summaries. We believe this is usually a reasonable assumption because notes are used at the point of care for this purpose. However, clinical notes, particularly in electronic systems, have been shown to often contain redundancies, incorporate outdated information, and omit important information.

There are several directions for future work. First, while our goal in this work was to demonstrate the utility of learning associations between structured health record data and clinical notes in a supervised learning framework, other modeling approaches should be explored. In addition, we considered clinical events and physiological time-series together as "structured data." Future work could investigate other methods for combining the two modalities of data and the relative utility of each in generating meaningful summaries.

Generating topic distributions of clinical notes could be useful in proposing potentially missing topics to care team members while they are writing a note. In addition, our approach is a first step towards a learning-based framework for generating clinical text that summarizes structured health record data. Future work could include generating candidate phrases corresponding to patient history. While our analysis is limited to the intensive care setting and to the structure and notes in MIMIC, our approach could similarly be used to generate topics summarizing longitudinal health record data in outpatient settings.

Acknowledgments

This work was funded in part by Wistron Corporation.

References

Vinett Arora, J Johnson, D Lovinger, HJ Humphrey, and DO Meltzer. Communication failures in patient sign-out and suggestions for improvement: a critical incident analysis. *BMJ Quality & Safety*, 14(6):401–407, 2005.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

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*, pages 69–78. ACM, 2015.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

François Chollet et al. Keras. https://keras.io, 2015.

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.

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 *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Marzyeh Ghassemi, Mike Wu, Michael C Hughes, Peter Szolovits, and Finale Doshi-Velez. Predicting intervention onset in the icu with switching state space models. *Proceedings of AMIA Summits on Translational Science*, 2017:82, 2017.

Ayelet Goldstein and Yuval Shahar. An automated knowledge-based textual summarization system for longitudinal, multivariate clinical data. *Journal of Biomedical Informatics*, 61: 159–175, 2016.

Jen J Gong, Tristan Naumann, Peter Szolovits, and John V Guttag. Predicting clinical outcomes across changing electronic health record systems. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1497–1505. ACM, 2017.

Ricardo Henao, James T. Lu, Joseph E. Lucas, Jeffrey Ferranti, and Lawrence Carin. Electronic health record analysis via deep poisson factor models. *Journal of Machine Learning Research*, 17(186):1–32, 2016. URL http://jmlr.org/papers/v17/15-429.html.

Jamie S Hirsch, Jessica S Tanenbaum, Sharon Lipsky Gorman, Connie Liu, Eric Schmitz, Dritan Hashorva, Artem Ervits, David Vawdrey, Marc Sturm, and Noémie Elhadad.

Harvest, a longitudinal patient record summarizer. *Journal of the American Medical Informatics Association*, 22(2):263–274, 2014.

Joyce C Ho, Joydeep Ghosh, Steve R Steinhubl, Walter F Stewart, Joshua C Denny, Bradley A Malin, and Jimeng Sun. Limestone: High-throughput candidate phenotype generation via tensor factorization. *Journal of Biomedical Informatics*, 52:199–211, 2014.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Vijay Huddar, Bapu Koundinya Desiraju, Vaibhav Rajan, Sakyajit Bhattacharya, Shourya Roy, and Chandan K Reddy. Predicting complications in critical care using heterogeneous clinical data. *IEEE Access*, 4:7988–8001, 2016.

Jim Hunter, Albert Gatt, François Portet, Ehud Reiter, and Somayajulu Sripada. Using natural language generation technology to improve information flows in intensive care units. In *ECAI*, pages 678–682, 2008.

Yohan Jo, Natasha Loghmanpour, and Carolyn Penstein Rosé. Time series analysis of nursing notes for mortality prediction via a state transition topic model. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 1171–1180. ACM, 2015.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, 2016.

Thomson Kuhn, Peter Basch, Michael Barr, and Thomas Yackel. Clinical Documentation in the 21st Century: Executive Summary of a Policy Position Paper From the American College of Physicians. *Annals of Internal Medicine*, 162(4):301–303, 2015.

Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzell. Learning to diagnose with LSTM recurrent neural networks. In *Proceedings of the International Conference on Learning Representations 2016*, 2016.

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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

Megan Monroe, Rongjian Lan, Hanseung Lee, Catherine Plaisant, and Ben Shneiderman. Temporal event sequence simplification. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2227–2236, 2013.

Rimma Pivovarov and Noémie Elhadad. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 22 (5):938–947, 2015.

Rimma Pivovarov, Adler J Perotte, Edouard Grave, John Angiolillo, Chris H Wiggins, and Noémie Elhadad. Learning probabilistic phenotypes from heterogeneous EHR data. *Journal of Biomedical Informatics*, 58:156–165, 2015.

Catherine Plaisant, Brett Milash, Anne Rose, Seth Widoff, and Ben Shneiderman. Lifelines: visualizing personal histories. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 221–227. ACM, 1996.

François Portet, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer, and Cindy Sykes. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173(7-8):789–816, 2009.

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

Hardeep Singh, Christiane Spitzmueller, Nancy J Petersen, Mona K Sawhney, and Dean F Sittig. Information overload and missed test results in electronic health record–based settings. *JAMA Internal Medicine*, 173(8):702–704, 2013.

Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding using deep networks. In *Proceedings of Machine Learning for Healthcare*, 2017.

Mike Wu, Marzyeh Ghassemi, 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*, 24(3):488–495, 2017.

Appendix A.

Table A1: Differences in length of stay, care units, admission type, and adverse outcome incidence between patients with and without physician, nursing, and general notes.

	Notes Missing	Notes Present
Number of patients	9171	6360
Mean LOS in ICU (days)	2.6	2.5
In-Hospital Mortality (%)	8.8	7.2
Intubation (%)	39.5	36.5
CCU (%)	12.4	13.0
CSRU (%)	17.1	16.7
MICU (%)	38.7	38.6
SICU (%)	18.8	18.1
TSICU (%)	13.0	13.7
Elective admission (%)	16.7	15.4
Emergency admission (%)	82.3	83.1
Urgent admission (%)	1.0	1.5

Appendix B.

Table B1: Top 10 tokens describing each topic.

Topic	Tokens
0	post, surgery, op, epidural, bladder, repair, iabp, stent, urology, pain
1	pleural, effusion, chest, tube, ct, effusions, fluid, drain, cxr, placement
2	fluid, na, stool, acidosis, diarrhea, diff, sodium, free, hyponatremia, cont
3	cancer, mass, ca, metastatic, lung, malignant, tumor, neoplasm, chemo, cell
4	skin, left, right, site, wound, groin, area, leg, impaired, intact
5	pain, abdominal, nausea, ct, vomiting, abd, ercp, zofran, iv, abdomen
6	lithium, morbid, myasthenia, suprapubic, mtx, girlfriend, atropine, cystitis, aureus, shocks
7	respiratory, pneumonia, pna, copd, aspiration, cxr, distress, bipap, sputum, nebs
8	code, continue, total, balance, rhythm, review, systems, labs, comments, prophylaxis
9	mental, status, altered, airway, delirium, cont, aspiration, agitation, agitated, risk
10	heparin, pe, ptt, started, dvt, gtt, pulmonary, transferred, cta, filter
11	impaired, problem, description, skin, enter, abscess, comments, integrity, tooth, clindamycin
12	right, left, ct, fractures, hematoma, injury, lobe, chest, posterior, thoracic
13	present, pulse, min, extremities, mmhg, current, regular, rhythm, insulin, chest
14	family, care, dnr, support, daughter, dni, son, comfort, morphine, social
15	etoh, abuse, ciwa, withdrawal, alcohol, pancreatitis, valium, scale, thiamine, seizures
16	liver, cirrhosis, lactulose, transplant, encephalopathy, ascites, hepatic, varices, sbp, albumin
17	seizure, sdh, dilantin, subdural, activity, neuro, seizures, brain, head, keppra
18	hct, bleeding, blood, stable, prbc, monitor, bleed, inr, cont, transfusion
19	afib, atrial, fibrillation, coumadin, rate, af, fib, po, metoprolol, amiodarone
20	gi, bleed, hct, bleeding, gib, egd, stable, gastrointestinal, protonix, upper
21	lasix, chf, diuresis, edema, failure, iv, heart, chronic, acute, goal
22	cath, cardiac, cad, heparin, chest, asa, nstemi, plavix, pain, disease
23	fever, temp, cont, wbc, cultures, sent, abx, cx, vanco, culture
24	neuro, commands, exam, extremities, eyes, pupils, checks, continue, noted, monitor
25	spontaneous, rr, min, set, vt, tube, ventilator, peep, mode, ve
26	arrest, cardiac, vt, icd, av, ccu, bradycardia, ep, rhythm, pacer
27	fx, fracture, fall, trauma, rib, collar, multiple, neck, injuries, pain
28	insulin, dm, diabetes, type, blood, gtt, scale, sliding, fs, bs
29	iv, order, total, extremities, rhythm, current, po, prn, fluid, balance
30	bed, oriented, oob, able, swallow, po, speech, chair, today, alert
31	present, normal, sounds, left, right, cardiovascular, respiratory, nose, pulse, absent
32	left, ct, head, hemorrhage, right, neuro, sbp, sah, stroke, sided
33	gtt, monitor, sbp, iv, bp, continue, remains, stable, noted, shift
34	neo, map, hypothermia, wean, pad, bair, hugger, temp, bypass, sfa
35	note, time, agree, section, protected, resident, present, saw, examined, services
36	intubated, sedation, vent, propofol, abg, extubation, sedated, fentanyl, wean, respiratory
37	hypotension, line, shock, sepsis, levophed, cvp, fluid, bp, pressors, map
38	cabg, artery, wires, coronary, bypass, temporary, graft, svg, avr, valve
39	likely, continue, pending, culture, negative, blood, cultures, infection, consider, cx
40	renal, failure, acute, hd, arf, chronic, cr, urine, bun, kidney
41	po, pain, denies, past, ed, prn, home, chest, prior, recent
42	valuables, transferred, rate, pmh, weight, heart, bp, total, sent, money
43	pain, control, chronic, acute, continue, prn, dilaudid, morphine, po, iv
44	abd, bowel, drainage, soft, output, urine, draining, bs, abdomen, ngt
45	ct, head, mri, status, mental, negative, osh, lp, spine, eeg
46	left, aortic, valve, right, normal, ventricular, mitral, systolic, stenosis, wall
47	ed, received, micu, bp, transferred, noted, iv, arrival, started, sent
48	sats, cough, nc, clear, face, mask, diminished, resp, bases, secretions
49	assessed, pulse, total, comments, left, right, balance, review, systems, labs

Appendix C.

Table C1: Number of units in LSTM layer (or fully-connected layer for model using static data) in final model configurations.

Model	**Hidden Layer Units**
Note Topic Distribution Prediction	
struct2note	256
notes2note	64
struct-notes2note	256
Outcome Prediction	
static	16
struct	32
notes	32
predicted notes	32