Exploratory Analysis of Electronic Health Records using Topic Modeling

Ivair Puerari[1], Denio Duarte[1], Guilherme Dal Bianco[1], Julyane Felipette Lima[1]

Federal University of Fronteira Sul
Campus Chapecó
Chapecó, Brazil
puerariivair@gmail.com
{duarte,guilherme.dalbianco,julyane.lima}@uffs.edu.br

Abstract. The rapid growth of electronic health record (EHR) systems brings an increase in available information about patients in hospitals. This massive amount of text information presents an opportunity to extract unknown information about medical history, medication, diseases, allergies, among others. Extracting the main topics that represent the subjects covered by a text collection can give valuable insights. To this end, approaches for topic modeling have been used to tackle such problems as information discovery and topic extraction with thematic information. In this context, this work presents an exploratory analysis of a collection of electronic health records from an intensive care unit (ICU). The collection is split into two sub-collections: discharged patients and patients who progressed to death. We apply an LDA-based approach to discover the latent topics from the collections. The analyses show that some topics are more recurrent in the deceased patients (the death collection), like renal diseases, and others are more recurrent in the discharge collection, for example, diabetes. The results of the analyses can be useful for improving intensive care services since the topics can be a guide to understanding the patterns in discharge and death situations.

Categories and Subject Descriptors: H.2 [**Database Management**]: Miscellaneous; H.3 [**Information Storage and Retrieval**]: Miscellaneous; I.7 [**Document and Text Processing**]: Miscellaneous

Keywords: Topic Modeling, Electronic Health Record, ICU, LDA, Discharge, Death

1. INTRODUCTION

Hospitals generate a massive amount of data about healthcare every day. Although a substantial part of the documents is still generated on paper, a major effort towards digitization is being developed [Meskó et al. 2017]. Electronic health records (EHR) are an example of the digital data generated by hospitals. EHR contain information about the medical history of the patients, *e.g.*, test results, diseases, prescriptions, and procedures. In most of the cases, EHR is collected in a non-systematic manner according to clinical need [Mihaela Coroiu et al. 2019], and generally, they are composed of documents written freely. The flexibility in writing EHR brings a side effect: it is hard to extract hidden information or structures from the documents.

The intensive care unit (ICU) is one of the most critical departments of a hospital. ICU provides support to the most severely ill patients in a hospital, and patients are monitored closely to assist in the early detection and correction of deterioration before it becomes fatal [Kane et al. 2007]. Accordingly, documents generated during the stay of the patient in the ICU contain essential information about her history, including the causes of the patient's discharge or death. The volume of documents generated from ICU patients is very high, for example, the ICU at the Beth Israel Deaconess Medical Center in Boston generated more than 60,000 documents from 2001 to 2012 [Johnson et al. 2016].

Generally, documents present several challenges for information extraction, including the same concepts described in different manners, typos, and high-dimensional data. Because of this, several types of research have been done in document information extraction (see [Piskorski and Yangarber 2013] for a comprehensive study). One promising approach to discover latent information from document collections is topic modeling [Blei et al. 2003]. The topics are discovered based on the co-occurrence of the words in a given document collection. Based on the discovered topics, documents are clustered into subjects that make it easier to understand the collection. According to the frequency, latent topics could be revealed – the topics emerge from the analysis of the original documents [Blei 2012]. EHR's documents are a rich source for topic modeling since several hidden subjects (topics) could be discovered from the documents. These topics can be used further to understand, for example, the causes of mortality.

Several works address the problem of discovering topics in EHR using topic modeling. For example, in [Chan et al. 2013], topic modeling was applied to mine cancer clinical notes to discover patterns in the notes and patient's underlying genetics. In the same way, Arnold et al. [2010] apply Latent Dirichlet Allocation (LDA) to compare patients' notes to the discovered topics. On the other hand, Perotte et al. [2011] use topic modeling to cluster discharge summaries into hierarchical concepts to better understand patterns in the records. Recently, Mihaela Coroiu et al. [2019] take advantage of topic modeling to detect the most relevant topics with the goal of assisting in patients' diagnoses. However, these works do not address the identification of the most frequent topics that focus on the cause of death and discharge of patients from ICU.

Accordingly, this work proposes an exploratory analysis of a collection of documents representing EHR get from *Medical Information Mart for Intensive Care III* (MIMIC III) [Johnson et al. 2016]. A probabilistic topic model approach is applied to extract the main topics from a subset of the MIMIC dataset. Valuable insights are produced by a survey of the topics previously identified. The following research questions guide our contributions:
RQ1: Do topic modeling approaches based on ICU EHR help to identify the causes of patients discharge or death?
RQ2: Do there exist topics (causes) related exclusively to death or discharge from ICU?
RQ3: Which are topics (causes) presented in both situations, *i.e.*, death and discharge.
RQ4: Which is the most prevalent body system associated with the causes of ICU discharge or death?

We intend to answer those questions to capture latent topics from EHR and conduct a survey with experts to evaluate the discovered topics. Moreover, we check the topics against the collection to identify the most predominant ones in the ICU. Note that we do not intend to predict the death or discharge of an ICU patient as Kim et al. [2011] and Awad et al. [2017] do; instead, we are interested in exploring the topics raised from ICU EHR.

The remainder of this paper is organized as follows. The following section presents some theoretical background to understand our proposal better. Sections 3 and 4 present how the experiments are conducted and the exploratory analyses. Section 5 discusses the work related to our proposal. Finally, Section 6 concludes this paper and presents some future work.

2. TOPIC MODELING

Topic modeling refers to a set of algorithms that aims to extract, given a collection of documents, the main topics that represent the subjects covered by the collection [Blei 2012; Steyvers and Griffiths 2007]. Topic modeling belongs to the unsupervised algorithms class, where input data do not have labels to categorize every example [Duarte and Ståhl 2019]. This brings some challenges for the creation and evaluation of the obtained models: the number of topics for a given collection, assessing the number of topics, and the most suitable evaluation metrics [Chang et al. 2009; Lau et al. 2014; Röder et al. 2015a].

2.1 Topics

Topics are derived from probabilistic word distributions in the input document collection. A set of words that by the relation of order, frequency, and semantics represent certain subjects (themes). Thus, through these relationships, it is possible to define a theme as a topic, that is, a probabilistic distribution of words with frequency and semantics that make sense within the topic's context.

Table I presents an example with four topics and their top-5 words with the respective probabilities of occurring in the topic (column $P(w)$). The table presents four possible topics from the MIMIC III collection. Notice that as there are no labels, the domain expert must define the semantics of each topic. For example, *Topic 2* should refer to the heart (or cardiovascular system), the word *valve* being the most likely to occur (*i.e.*, 3.1%).

Table I: Top-5 words of four topics in MIMIC III.

Topic 1		**Topic 2**		**Topic 3**		**Topic 4**	
P(w)	word	P(w)	word	P(w)	word	P(w)	word
0.017	skin	0.031	valve	0.020	liver	0.009	arrest
0.013	drain	0.026	aortica	0.016	bleed	0.008	transfer
0.012	wound	0.021	ventricular	0.015	renal	0.007	daily
0.011	draining	0.017	mitral	0.011	cirrhosi	0.007	comfort
0.011	fractur	0.014	leaflet	0.009	hepatic	0.007	pulse

2.2 Documents

Topic modeling is based on the idea that documents are mixtures of topics, *i.e.*, documents display multiple topics [Steyvers and Griffiths 2007; Blei 2012]. Thus, documents can be generated from different distributions on topics. A document can be defined as a sequence of words $\mathbf{w}=(w_1, w_2, \ldots, w_n)$, where n is the number of words in $\mathbf{w}$. Similarly, a corpus (or collection) is a set of m documents $D=\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m\}$. Moreover, a document can be any text content, *e.g.*, an article or comment on a social network.

In topic modeling, most approaches consider the document as a bag-of-words, that is, the order of the words in the document does not matter. Moreover, a pre-processing must be performed on the collection of documents to prepare it for extracting the topics. The pre-processing phase can be composed of the following steps[Steyvers and Griffiths 2007]: (*i*) removal of stop-words, *i.e.*, removing spurious words from the collection, (*ii*) tokenization, *i.e.*, transforming the collection into a list of words, (*iii*) stemming, *i.e.*, reducing the words to their root form, and (*iv*) lemmatizing, *i.e.*, grouping together the inflected forms of a word.

2.3 Approaches for Topic Modeling

Topics bring together document collections into groups (or clusters). There are three main approaches to discovering topics from document collection [Xie and Xing 2013]: clustering, matrix factorization, and LDA-based. Clustering methods include k-means (*e.g.* [Alhawarat and Hegazi 2018]), Spectral (*e.g.* [Huang et al. 2013]) and Hierarchical clustering (*e.g.* [Fung et al. 2003]). Non-Negative Matrix Factorization (NMF) is the most popular method for matrix factorization, and several approaches rely on NMF to discover topics from EHRs: [Kuang et al. 2015], [Luo et al. 2017], and [Zhao et al. 2019]. On the other hand, LDA-like approaches have been widely used to discovered hidden topics in document collections. The success of LDA-like approaches is due to the fact that they have been

conceived to extract topics. Chen et al. [2015] claims that using LDA to learn phenotypic topics exhibits high consistency. Moreover, some works, as [Chen et al. 2016], [Lu et al. 2016], [Suri and Roy 2017], and [M'sik and Casablanca 2020], point out that LDA-like approaches perform as well as or better than NMF approaches.

For our exploratory analysis and based on the studied literature, we decide to apply an LDA-like approach. The LDA [Blei et al. 2003] is one of the most used probabilistic modeling algorithms to extract topics from collections of documents. It is characterized by initially assigning probabilities to the words in the dictionary extracted from the collection. Distribution is done using Dirichlet's multivariate discrete distribution family.

Figure 1 represents pictorially the LDA model [Blei et al. 2003]. The plates represent iterations: the outer one represents the documents, and the inner one represents the repeated choice of topics and words within a document. Moreover, assuming LDA as a generative process, Figure 1 can be explained as follows:

(1) For each document w in a corpus $\mathcal{D}$:
   (a) Choose $N \sim \text{Poisson}(\xi)$
   (b) Choose $\Theta \sim \text{Dir}(\alpha)$
   (c) For each of the N words w_n:
      i. Choose a topic $z_n \sim \text{Multinomial}(\Theta)$
      ii. Choose a word w_n from $p(w_n \mid z_n\beta)$, multinomial probability conditioned on the topic z_n

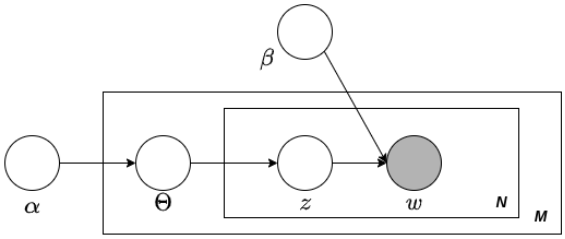


Fig. 1: Graphical model representation of LDA.

The hyperparameter β is the prior observation count on the number of times words are sampled from a topic before any word from the corpus is observed, higher β, more words are associated with a given topic. The hyperparameter α plays the same role but regarding the documents. Note also that LDA considers that documents exhibit multiple topics since a document, for example, about politics, can discuss economy and corruption. However, each topic associated with documents has a different probability, and the sum of all topics' probability associated with a given document is equal to one.

2.4 Metrics

Metrics are used to measure and evaluate models in machine learning. In topic modeling, as in any unsupervised learning model, assessing models is challenging because the datasets do not have labels to check the consistency of the results. The evaluation could be done by humans; however, it is an onerous task [Röder et al. 2015b].

In this context, Röder et al. [2015b] present a study comparing various coherence metrics for topic models. The study aimed to find which metric is the closest to the human assessment of the topics. The metric most correlated with human perception was C_v, which is based on the standardized version of *PMI* (Pointwise Mutual Information) (see Equation 1) in which the result is fixed in the range $[-1, 1]$, where 1 indicates complete occurrence between the words (w_i and w_j) and -1 no occurrence. Also, a sliding window is used to calculate the occurrence of words. For example, given a set of words

$C = \{w_1, w_2, w_3, w_4, w_5, w_6\}$, one sliding window of size three may contain the window $W_1 = \{w_2, w_3, w_4\}$ and slide to $W_2 = \{w_3, w_4, w_5\}$.

$$PMI(w_i, w_j) = \log\left(\frac{P(w_i, w_j) + \epsilon}{P(w_i) \cdot P(w_j)}\right) \quad (1)$$

Due to the C_v metric's excellent performance regarding the correlation with human perception, we use it in our experiments to find the best hyperparameters (*e.g.* number of topics) to apply in the MIMIC III collection.

3. EXPERIMENT SETUP

In this section, we present the configuration used in our experimental environment. More specifically, we describe where our exploratory analysis falls, the collections' characteristics, pre-processing, and the hyperparameters' configuration.

3.1 Experiment design

We extract from the EHR data the answers to our research questions follow a particular study design. The major groups of study designs are Experimental and Observational [Yadav et al. 2018]. In the former, the researcher intervenes by changing the course of the experiments and observes the resultant outcome. In the latter the researcher does not interfere with the outcome's result, and it is the most commonly used. Figure 2 shows the hierarchy of the study design regarding data mining in the EHR proposed by [Yadav et al. 2018].

Our exploratory analysis falls into Descriptive study design since it is Observational, and we do not have any outcome of interest (unsupervised learning), i.e., the study is designed to analyze the distribution of variables, without regard to an outcome [Yadav et al. 2018]. As we do not consider the time in our analysis, the study is also atemporal.

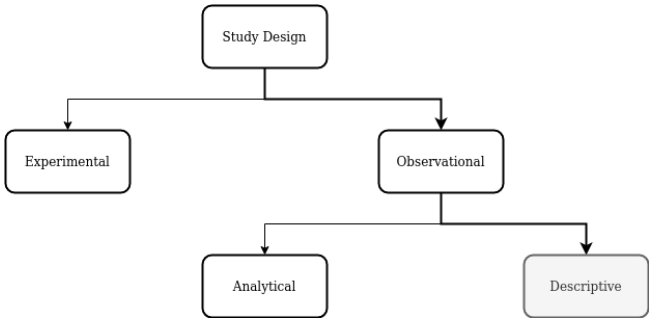


Fig. 2: Study design classification hierarchy.

3.2 Experiment implementation

We implement our experiment using the language Python and Gensim LDA implementation [Řehůřek and Sojka 2010]. The experiments ran on a Core 10 Intel(R) Xeon(R) CPU E7–4850 @ 2.00GHz 1064.444 Mhz, 10 Cores, 80 CPU, 126 GB of memory, and 6 terabytes of hard disk running Linux server.

3.3 Data Acquisition

The collection of documents with electronic health records comes from the ICU sector, specifically taken from the Medical Information Mart for Intensive Care III (MIMIC III) [1] database. MIMIC is a large database, freely available under request, which includes health-related data from 53,423 different hospital admissions. There were 38,597 adult patients aged over 16 years and 7,870 newborn patients. These patients were in ICU at the Beth Israel Deaconess Medical Center in Boston, Massachusetts, between 2001 and 2012.

MIMIC III is composed of 26 relational tables. The tables are listed by identifiers that usually contain the suffix ID as the attribute, *e.g.*, `subject_id` is the unique identifier for table `PATIENT`. The tables used in our work are briefly presented in the following. Table `ADMISSIONS` stores the hospitalization data for patients. Each hospitalization is associated with a unique identifier, represented by the attribute `hadm_id`. In `ADMISSIONS`, there is the `diagnosis` attribute, which describes the preliminary diagnosis in free text for the patient on hospital admission. The diagnosis is usually assigned by the doctor on duty and does not use a systematic ontology, that is, they describe signs or symptoms and possible diagnosis. Finally, the `hospital_expire_flag` attribute that indicates whether the patient died or was discharged during hospitalization.

Five of the database tables correspond to the MIMIC III data dictionary: `D_CPT`, `D_ICD_DIAGNOSES`, `D_ICD_PROCEDURES`, `D_ITEMS`, and `D_LABITEMS`. These tables are used as a reference for other tables (via foreign keys) to obtain descriptions of procedure codes, items, among others. Table `D_ICD_DIAGNOSES` stores the codes of the International Classification of Diseases Version 9 (ICD-9) for diagnosis, and it is composed of the attributes `short_title` and `long_title`, which provide two types of description for the code. Table `DIAGNOSES_ICD` stores the patients' diagnoses.

Table `NOTEEVENTS` keeps all notes for the patient. The `category` attribute defines the type of note, for example, the value "Discharge" indicates that the note is for a discharged patient. The `text` attribute contains the observation itself, in free text. In the observations, all actions and results about the patient in a given event are described. `procedures_icd` contains the procedures performed during the patient's hospitalization. The ICD-9 code identifies a specified procedure, which can be associated with the table `D_ICD_PROCEDURES` to determine which procedure is recorded for the patient. The `D_ICD_PROCEDURES` contains the attributes `short_title` and `long_title` that store the performed procedure code.

3.4 Collections Discharge and Death

The original database was imported into the Relational DBMS PostgreSQL. From the created database, the documents for the discharge and death collections were built as follows: (*i*) the attribute `diagnosis` (Table `ADMISSION`) contained the original text of the hospitalization, (*ii*) the attribute `long_title` (Table `D_ICD_DIAGNOSIS`) contained all diagnoses made during the hospitalization, (*iii*) the attribute `long_title` (Table `D_ICD_PROCEDURES`) represents the procedures performed; and, finally, (*iv*) the event notes during hospitalization were extracted from the `text` attribute of the table `NOTEEVENTS`.

The queries performed to extract textual data from the database and then create the collections are shown in Figure 3. To select only hospitalizations that evolved to death or discharge, the filter used in the query is the attribute `hospital_expire_flag`, in which the value 1 represents deaths and 0 discharges. Note that four queries build the collections and `hadm_id` is used to group the admissions. All occurrences of one admission were used to create a single document for that admission.

The collection of deaths consists of 6,051 documents, with an average of 10,931 words per document. The discharge collection has 53,954 documents, with an average of 7,174 words per document. Most

[1] `mimic.physionet.org`

```
SELECT hadm_id, diagnosis
FROM admissions
WHERE hospital_expire_flag = {}
```

```
SELECT a.hadm_id, ne.text
FROM admissions a JOIN noteevents ne
ON ne.hadm_id = a.hadm_id
WHERE a.hospital_expire_flag = {}
```

```
SELECT a.hadm_id, dip.long_title
FROM admissions a JOIN procedures_icd pi
ON pi.hadm_id = a.hadm_id
JOIN d_icd_procedures dip
ON dip.icd9_code = pi.icd9_code
WHERE a.hospital_expire_flag = {}
```

```
SELECT a.hadm_id, did.long_title
FROM admissions a JOIN diagnoses_icd di
ON di.hadm_id = a.hadm_id
JOIN d_icd_diagnoses did
ON did.icd9_code = di.icd9_code
WHERE a.hospital_expire_flag = {}
```

Fig. 3: Queries that build the collections.

of the documents belong to the discharge collection, representing almost 90% of the documents of the two collections. Every document in the collections represents the medical record of a patient who was admitted to the ICU. The medical record of discharge collection is formed, on average, by the concatenation of 30 documents, while the death collection is formed on average of 43 documents.

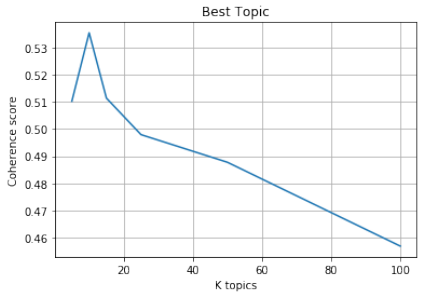
3.5 Pre-processing

After creating the collections, a pre-processing step was performed. Pre-processing is necessary because a document is mostly composed of notes on events held during hospitalization in a free text format. Table II(A) shows an original EHR from MIMIC II, and it can be seen that there is no pattern in writing it. This document excerpt consists of many punctuation marks, special symbols, uppercase and lowercase words, conjunctions, and articles.

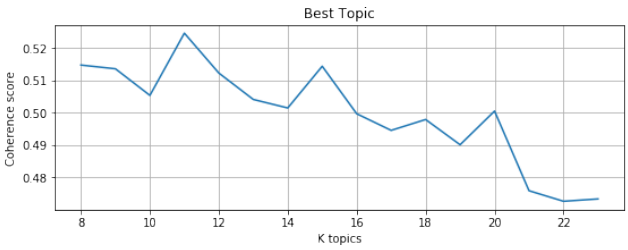
Table II: An original EHR and its post-processed version.

A - Original EHR document:
STATUS EPILEPTICUS Grand mal status Hodgkin's disease, unspecified type, unspecified site, extranodal and solid organ sites Postinflammatory pulmonary fibrosis Pneumonia, organism unspecified Unspecified essential hypertension Macular degeneration (senile), unspecified Antineoplastic and immunosuppressive drugs causing adverse effects in therapeutic use Other generalized ischemic cerebrovascular disease Open biopsy of brain Continuous invasive mechanical ventilation for 96 consecutive hours or more Spinal tap Venous catheterization, not elsewhere classified Enteral infusion of concentrated nutritional substances Venous catheterization, not elsewhere classified Admission Date: [**2108-8-22**] Discharge Date: [**2108-8-30**] Date of Birth: [**2036-5-17**] Sex: M Service: Neurology HISTORY OF PRESENT ILLNESS:
B - Same document after pre-processing:
statu epilepticu grand statu hodgkin diseas unspecifi type unspecifi site extranod solid organ site postinflammatori pulmonari fibrosi pneumonia organ unspecifi unspecifi essenti hypertens macular degener senil unspecifi antineoplast immunosuppress drug caus advers effect therapeut gener ischem cerebrovascular diseas open biopsi brain continu invas mechan ventil consecut hour spinal venou catheter elsewher classifi enter infus concentr nutrit substanc venou catheter elsewher classifi admiss discharg birth servic neurolog histori present known lastnam yearold gentleman histori

As previously presented (Section 2.2), pre-processing consists of removing punctuation and special symbols; transformation of words to lowercase; lemmatization; stemming; removal of digits; removal of words with less than three letters; and removing *stopwords*. Pre-processing causes a significant decrease in the size of the document. Again, Table II(B) shows a post-processed version of the EHR in which the result of pre-processing on the document is noticeable. Note that there are no punctuation marks, all words are lowercased, and some words are lemmatized, *e.g. unspecified* becomes *unspecifi* and *extranodal* becomes *extranod*.



(a) Topics coherence versus number of topics.



(b) Topic coherence for experimenting K and α.

Fig. 4: Results of experiments to find the best K and the best α.

3.6 Hyperparameters Configuration

The Gensim LDA implementation needs at least two hyperparameters to build an LDA model: number of topics (K) and α, which represents the strength of the distribution of documents by topics. The higher the α is, the more topics are associated per document.

Experiments were carried out to find the best value for K and α. The C_v metric was used to evaluate the best combination for these two hyperparameters. We first ran six experiments to check best value for K: 5, 10, 15, 25, 50, and 100. For each topic result, the coherence metric was applied. Figure 4a presents a plot of the performance of each value for the number of topics. Looking at the figure, we see that the coherence decreases when $K > 15$, becoming less than 0.50 from 22 topics.

Another experiment was conducted to find the final best K. The values for K were [8, 9,10,..., 23] that presented better coherence value in the previous experiments. For each K, we set α as [0.01, 0.31, 0.61, 0.90999, symmetric, asymmetric] to find the best one. The values symmetric and asymmetric are set automatically as $1/K$ and $1/(\sqrt{K}+1)$, respectively. Figure 4b shows the best results of the combinations of K and α. Note that the best value for K is 11, so K is set to 11, and α is set to 0.01 for the final experiments.

Also, the dictionary size was defined at this stage. Through the previous experiments, the dictionary was configured to have a maximum of 2,000 words, in such way that words which appeared in under 10% or over 80% of the documents were excluded. The exclusion of less and more frequent words reduced the sizes of the discharge and death collections to 1,945 and 1,662 words, respectively.

4. EXPERIMENT

Here, we conduct a set of exploratory experiments applying the hyperparameters previously identified in death and discharge collections. After discovering the 11 topics for each collection, the topics were organized into top-10 words, that is, the first ten words most likely to occur in the topic. In this and the following sections, we present only the top-5 topics, *i.e.*, the five topics with more documents associated. We refer the readers to APPENDIX A for the list of all 11 topics.

As presented previously, documents are mixtures of topics, and we check the association between documents and topics. As we want to pick only documents with just one strong topic, we limit the probability of a topic to be associated with a document to 50%. Remember that the sum of the probabilities concerning the topics in a document is equal to one. Figure 5 shows the ranking produced by this association. Notice that there are topics that are stronger than others. In the death collection, for example, Topic 5 is the most prominent, but in the discharge collection, Topic 1 is the one that occurs most in the collection. Note also that the association of the document to topics is smoother in the discharge collection than in the death collection, meaning that in the death collection, some

topics are more prominent than others.

For the purpose of clarity, Table III and IV present only the top-5 topics of the 11 in the death and discharge collections, respectively (all the topics can be found in APPENDIX A). That is, the five topics most likely to occur in the collections. Topics are organized in the table from left to right, where the leftmost is the most frequent topic. For example, Topics 5 and 1 are the most frequent in the collections of death and discharge, respectively. In parenthesis, we have indicated the frequency of every topic. Section 4.1 presents an evaluation of the discovered topics.

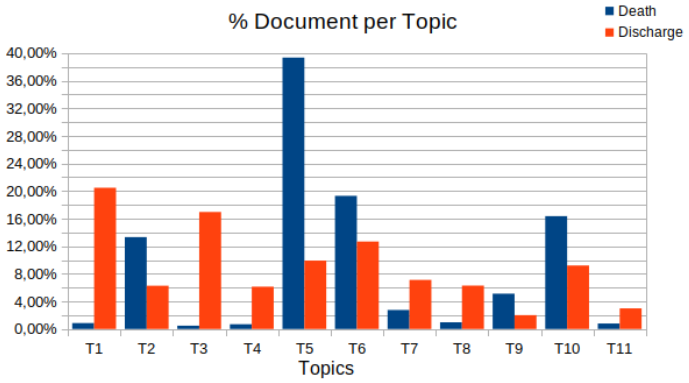


Fig. 5: Percentage of documents per topics in both collection.

Note that the range of probabilities of the words in Topic 5 from the discharge collection is very small. That means there is no strong word in this topic; for example, we can see this in Topic 10. It seems that Topic 5 is related to at least three subjects: respiratory system, cardiac system, and renal system. Section 4.2 explores in detail the relationship between topics and human body systems.

Table III: The top-5 topics and their top-10 words from death collection.

Topic 5 (39.36%)		**Topic 6 (19.31%)**		**Topic 10 (16.37%)**		**Topic 2 (13.32%)**		**Topic 9 (5.12%)**	
P	word	P	word	P	word	P	word	P	word
0.014	respiratory	0.030	hemorrhage	0.014	pleural	0.031	valve	0.025	contrast
0.011	ventilation	0.025	head	0.013	pneumonia	0.026	aortica	0.013	abdomen
0.008	wean	0.011	contrast	0.010	unchanged	0.021	ventricular	0.012	catheter
0.008	neuro	0.011	neuro	0.010	opacities	0.017	mitral	0.012	liver
0.008	secretion	0.010	intubated	0.010	interval	0.014	leaflet	0.011	identifier
0.007	intubated	0.010	frontal	0.009	lobe	0.013	systole	0.010	within
0.007	thick	0.010	seizure	0.008	pneumothorax	0.012	wall	0.010	vein
0.007	suction	0.009	cerebral	0.007	upper	0.011	mild	0.009	abdomin
0.007	shift	0.009	subarachnoid	0.007	comparison	0.010	regurgit	0.009	pelvis
0.006	urin	0.009	ventricles	0.006	worsen	0.010	mildline	0.008	evidence

4.1 Topics Evaluation

We conduct a qualitative evaluation with experts (as done by [Bai et al. 2017]) to validate and cluster the discovered topics. Eight senior students from the Nursing Course at the University of Santa Catarina State (UDESC), who have already performed theoretical-practical activities at ICU, were asked to answer a survey to label the topics. First, we presented how the collections were built and how the topics were extracted. In the discussion, we gave the necessary information about our work, presenting a summary. The experts do not have access to the documents for the sake of confidentiality

Table IV: The top-5 topics and their top-10 words from discharge collection.

Topic 1 (20.48%)		**Topic 3 (19.96%)**		**Topic 6 (12.68%)**		**Topic 5 (9.91%)**		**Topic 10 (9.20%)**	
P	word	P	word	P	word	P	word	P	word
0.030	valve	0.016	sepsis	0.022	head	0.009	acute	0.054	feed
0.024	aortica	0.013	baby	0.022	fractur	0.008	urine	0.022	active
0.019	arteria	0.013	murmur	0.019	hemorrhage	0.008	pulse	0.022	stool
0.018	ventricular	0.013	nicu	0.018	contrast	0.007	bleed	0.015	respiratory
0.016	mitral	0.013	newborn	0.010	neuro	0.007	respiratory	0.014	murmur
0.013	leaflet	0.013	feed	0.010	arteria	0.007	fluid	0.011	retract
0.012	coronary	0.011	active	0.010	radiolog	0.007	rhythm	0.011	cpap
0.011	cardiac	0.011	born	0.008	evidence	0.007	renal	0.010	benign
0.011	systole	0.010	week	0.008	mass	0.007	chronic	0.010	neonatolog
0.010	wall	0.010	screen	0.008	hematona	0.006	stool	0.010	week

following the *Internet Research: Ethical Guidelines 3.0 Association of Internet Researchers* (`aior.org`). Afterward, they answered the survey.

The survey was divided into two sections, one for each collection (i.e., discharge and death). Each section presented the 11 topics with their top-10 words. The respondents should answer the following question "Which body system is affected or procedure or complication, or treatment corresponds to this set of words?". This question was elaborated from the authors' perspective and knowledge about the subject. Charmaz [2009] states that, although the researcher does not want to bias the exploratory analysis, her beliefs and knowledge about the subject can be part of the whole process, *i.e.*, researchers can never eliminate their bias or beliefs from research.

The multiple choices for each section were defined based on the topics (exclusivity choices for the collections are in italic):

—Discharge: Cardiac System, Diabetes, *Postoperative*, *Pancreas*, *Liver*, Hepatic System, Hepatic/Pancreas System, Sepsis, Prematurity, Surgery, Respiratory System, *Cleft Lip*, *Hemorrhage*, Cardiovascular System, Renal System, Neurological System, Cranioencephalic Trauma (TBI), Digestive System, Respiratory and Cardiovascular Systems, and *Cancer*.

—Death: Respiratory System, Sepsis, *Integumentary system*, *Trauma/Burn Injuries*, *Politrauma*, Cardiovascular System, Hepatic System, Urinary/Renal System, Renal System, Cranioencephalic Trauma (TBI), Neurological System, *Respiratory System associated to Neurological problems*, Digestive System, *Venous Catheterization*, *Gastrointestinal System*, Surgery, and *Pleural Effusion*.

The results of the survey are shown in Tables V and VI for discharge and death collections, respectively. The choices of the subjects were mainly related to body systems. There are 20 different labels for the discharge collection, while the death collection presented 17 different subjects. On average, two different subjects were listed for each topic in the discharge collection, whereas the death collection presents an average of 3 subjects for each topic.

In general, the results obtained are considered consistent since they were related to the objective of this work. Thus, all topics extracted from the subjects selected by the survey could to be used to define the subject (label) for each topic. Next, we present a detailed discussion of such results.

4.2 Results Discussion

As shown in Table V (discharge collection), Topics 1, 4, 7, 9, and 11 are related to just two labels: "Cardiac System" and "Respiratory System" (four times), respectively. It means that four out of 11 topics are about the respiratory system, that is, it is the most predominant topic in the discharge collection (**RQ2** and **RQ4**). The study carried out in [da Silva et al. 2020] supports this claim since up to 79.2% of patients in the ICU are submitted to mechanical ventilation. On the other hand,

Table V: Topics and subjects association (Discharge Collection).

	Topics										
Subjects	**01**	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**	**11**
Cardiac System	8										
Diabetes		3									
Postoperative		1									
Pancreas		2									
Liver		1									
Hepatic/Pancreas System		1									
Sepsis			3								
Prematurity			4							2	
Surgery			1								
Respiratory System				8			8		8	2	8
Hemorrhage					3						
Cardiovascular System					1						
Renal System					4			2			
Neurological System						5					
Cranioencephalic Trauma (TBI)						3					
Hepatic System								5			
Digestive System								1			
Respiratory and Cardiovascular Systems										2	
Cancer										2	

Table VI: Topics and subjects association (Death Collection).

	Topics										
Subjects	**01**	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**	**11**
Respiratory System	2				6		5			6	
Sepsis	3		1								3
Integumentary system	1										
Trauma/Burn Injuries	1										
Politrauma	1										
Cardiovascular System		8		7			1	3			1
Hepatic System			5						3		
Urinary/Renal System											1
Renal System			1				2	5	1	1	3
Cranioencephalic Trauma (TBI)				1							
Neurological System					1	8					
Respiratory System associated to Neurological problems					1						
Gastrointestinal System									2		
Venous Catheterization									1		
Gastrointestinal Hepatic System									1		
Surgery			1								
Pleural Effusion										1	

Topics 2, 3, 5, 8, and 10 are associated with three or more subjects. Note also that Topic 6 can be related to the "Neurological System" since TBI can also be related to this system.

Regarding the death collection (see Table VI), the most assertive topics are 2 and 6, "Cardiac" and "Neurological" systems, respectively (**RQ2** and **RQ4**). "Cardiac System" is chosen partially for Topics 4, 7, 8, and 11, meaning that it is a recurrent subject in the death and discharge collections (**RQ3**). The same reasoning can be used regarding the "Respiratory System". "Sepsis" is also a subject recurrent in the death collection (*i.e.*, related to three topics). Topics 2 and 6 also belong to the top-5 topics from the death collection. The other three top-5 topics (*i.e.*, 5, 9, and 10) are very mixed regarding the subjects.

Looking through both collections' topics, we see that the subjects are evenly distributed in the topics, but "Respiratory Systems" appears in five topics. However, there are strong subjects like "Respiratory System", "Cardiovascular System", "Sepsis", and "Renal System" in the death collections.

We arrange subjects in similar groups. This grouping can help to identify the most prevalent subjects by collection. We propose the following grouping:

—Heart Related
—Discharge: Cardiac System, Cardiovascular System, and Respiratory and Cardiovascular Systems

—Death: Cardiovascular System and Venous Catheterization
—Lung Related
—Discharge: Respiratory System and Respiratory and Cardiovascular Systems
—Death: Respiratory System, Respiratory System associated to Neurological problems, and Pleural Effusion
—Digestive Related
—Discharge: Pancreas, Liver, Hepatic System, Hepatic/Pancreas System, and Digestive System
—Death: Gastrointestinal System, Gastrointestinal Hepatic System, Hepatic System, and Gastrointestinal Hepatic System
—Sepsis Related: Sepsis
—Renal Related
—Discharge: Renal System
—Death: Urinary/Renal System and Renal System
—Neurological
—Discharge: Neurological System
—Death: Neurological System and Respiratory System associated to Neurological problems
—Others
—Discharge: Prematurity, Surgery, Diabetes, Hemorrhage, Postoperative, Cranioencephalic Trauma (TBI), and Cancer
—Death: Integumentary system, Trauma/Burn Injuries, Politrauma, Cranioencephalic Trauma (TBI), and Surgery

Figure 6 presents two pie charts showing the distribution of subjects by groups: discharge and death. Without considering Others, the group related to the lung is widespread in both collections (seven and six topics in the discharge and death collections, respectively) (**RQ3**). In the death collection, renal disease problems are the most frequent (seven topics). That means that kidney disease is likely to lead to death in ICU (**RQ2** and **RQ4**). The kidney seems a less important organ of the human body than the heart and lung; however, studies show that renal system failure is rising as a cause of death in the ICU [Chertow et al. 2006; Dare et al. 2017] (**RQ4**). However, in the discharge collection, it is much less frequent. Topics related to heart problems are twice as frequent in death collections as in the discharge collection (six against three) – another interesting point: the groups in death collections are well distributed along with the topics. That is, it is more difficult to identify a predominant group in that collection (**RQ2** and **RQ3**).

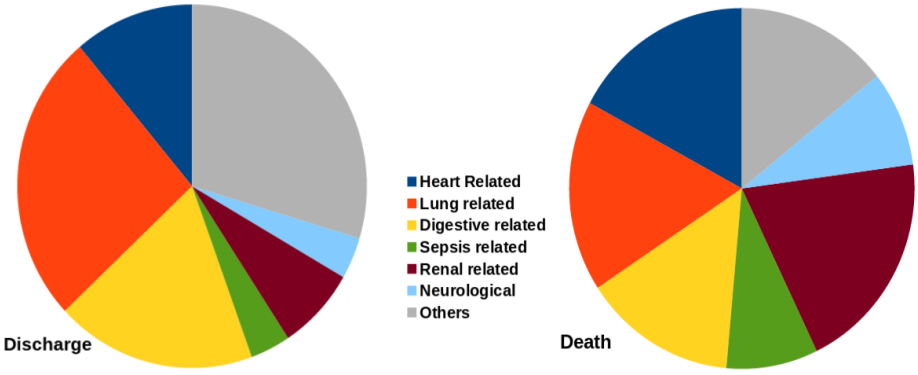


Fig. 6: Number of topics associated to the subject groups.

If we analyze the subjects associated with the topics (Tables V and VI), we notice that Diabetes and Prematurity are not associated with topics from the death collection. Indeed, studies show that

both have a positive evolution in most cases in ICU. A recent study that analyzed neonatal mortality by analyzing medical records in two consecutive years identified a mortality rate of approximately 12.5% [Sena et al. 2020]. Diabetes is a disease that affects a considerable global population, and it is estimated that approximately half of the affected people are unaware of their condition. Diabetes appears in the discharge collection since it affects a considerable percentage of people and is associated with several other conditions that can lead to hospitalization [Guariguata et al. 2014]. However, Koye et al. [2018] claims that kidney disease is significantly related to Diabetes, *i.e.*, Diabetes does not appear in the death collection, but Renal related topics are predominant.

To summarize, the analyses show that tacit knowledge can be confirmed by looking through the topics, and that confirms **RQ1**, *i.e.*, topic modeling approaches help to identify and understand the causes of patients discharge or death based on ICU EHR.

5. RELATED WORK

There is a vast diversity of works on topic modeling using electronic medical records in the literature. Most of them use topic modeling as a pre-processing step to apply any machine learning supervised algorithm further. Here, the focus is on approaches related to our proposal (for an overview, see [Chen et al. 2017] and [Jelodar et al. 2019]).

Some works are similar to ours if we consider the classification presented in Figure 2, *i.e.*, they are descriptive. For example, Doshi-Velez et al. [2014], Kalankesh et al. [2013], Gotz et al. [2011], and Roque et al. [2011] apply clustering algorithms to group EHRs. The works proposed in [Kalankesh et al. 2013] and [Doshi-Velez et al. 2014] explore hierarchical clustering to investigate patterns in patients like comorbidity and patient stratification for the discovery of overlapping genes and co-occurring diseases for patients diagnosed with autism spectrum disorders. On the other hand, [Gotz et al. 2011] and [Roque et al. 2011] apply clustering approaches to explore clusters of similar patients. Roque et al. [2011] pre-process EHRs to reduce the dimensionality of the collection and improve the final visualization. Yet our goal is not to group patients by disease or diagnostic but to identify the causes of death or discharge of patients in ICU.

Regarding the use of LDA-based approaches to extract topics or clusters from a document collection, Ding Cheng et al. [2014] conducted a study using the LDA to cluster patients in diagnosis-groups, represented by ICD-9 codes, for identifying comorbidity. The authors are interested in discovering diagnosis code groups through topic modeling. They do not check the relevant associations; nor evaluate the coherence of the topics.

The approaches proposed in [Lehman et al. 2012; Lehman et al. 2014] use Dirichlet Hierarchical Process (HDP) to extract topics from EHR. HDP is an extension of LDA that discovers the appropriate number of topics of a collection automatically. Lehman et al. [2012] combined the discovered topics structure of Unified Medical Language System (UMLS) clinical concepts extracted from the first 24-hour ICU nursing notes with physiologic data for risk stratification of in-hospital mortality. However, Lehman et al. [2014] focused on discovering clinical topics in discharge summaries that are predictive of post-hospital discharge mortality. They discovered useful categories such as *on ventilator*, *post-cardiac surgery*, *trauma*, and *pulmonary disorders*.

Mihaela Coroiu et al. [2019] propose a method to assist doctors in establishing patients' diagnoses based on the analysis of medical records. The approach detects most relevant topics (using LDA) from medical records. The discovered topics allow the physicians to make decisions with much more information than is usually available, that is, the topics serve as a guide to the diagnosis.

The work of Zhang et al. [2017] proposes a new model called the Survival Topic Model (SVTM). SVTM generates patient topics using data such as measurements, notes, and death/discharge information about trauma patients to create a new data set. Based on this dataset, the model predicts

the probability of death/discharge as a function of time. The idea is to distribute patients by disease conditions that are defined by topics. They discovered that injuries are highly correlated with each other. This is because these correlated injuries are located near to one another in the human body; thus, patients may have these multiple injuries simultaneously.

In the study conducted by Valenti et al. [2019], topic modeling is used to infer the emotional state of people living with Parkinson's disease. For this purpose, two models were used: LDA and Linguistic Inquiry and Word Count (LIWC), in order to evaluate which one is the best for determining the emotional state of the patients. The collection of documents comes from interviews during a randomized clinical trial that includes open-ended questions about events in daily life in the recent past, which participants had experienced and should rate them as frustrating or pleasant. LDA and LIWC were used to extract attributes from documents. The attributes were used in a classifier to create a model to identify a text as pleasant or not. The results obtained showed that LDA is suitable when documents are from medium to a small size; otherwise, LIWC is indicated.

Our work differs from the above since we are interested in identifying the prominent topics in a document collection extract from MIMIC III. We want to discuss the latent topics from EHR regarding discharge and death situations. We show that topic modeling can be a powerful tool for professional health care in their daily tasks.

6. CONCLUSION

In this work, we conduct an exploratory analysis of electronic health records (EHR) from ICU using a publicly available database. EHRs were divided into two collections of documents: discharge and death collections. The analysis was done grouping the documents by topics based on the LDA approach. Each collection was grouped into 11 topics, and, based on the top-10 words of every topic, we suggest which subjects better describe the topics. We propose five research questions to which our exploratory analysis has responded.

The discharge collection is mainly characterized by subjects such as the respiratory system, renal system, neurological system, prematurity, and cardiac system. On the other side, Death collection is characterized by the hepatic system, cardiovascular system, neurological system, and respiratory system.

We claim that these analyses can be very useful for health professionals since discovering subjects related to EHR helps to improve the care protocols. That information allows the professionals to assess and, in a certain way, to evaluate the actions and plans. A study carried out by Denaxas et al. [2016] indicates that health data must be made available for research to provide models, solutions, and tools to help the health system in general.

In future work, we intend to analyze the collections by slices of time. Using the slices, we can check the evolution of the topics, *e.g.*, the evolution of the diseases, and which subjects are more or less volatile in time. Another direction is to use Hierarchical LDA to identify the best number of topics in a collection automatically as well as to apply clustering-based approaches.

REFERENCES

Alhawarat, M. and Hegazi, M. Revisiting k-means and topic modeling, a comparison study to cluster arabic documents. *IEEE Access* vol. 6, pp. 42740–42749, 2018.

Arnold, C. W., El-Saden, S. M., Bui, A. A., and Taira, R. Clinical case-based retrieval using latent topic analysis. In *AMIA annual symposium*. Vol. 2010. American Medical Informatics Association, pp. 26, 2010.

Awad, A., Bader-El-Den, M., McNicholas, J., and Briggs, J. Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach. *International journal of medical informatics* vol. 108, pp. 185–195, 2017.

Bai, T., Chanda, A. K., Egleston, B. L., and Vucetic, S. Joint learning of representations of medical concepts and words from ehr data. In *2017 IEEE BIBM*. pp. 764–769, 2017.

Blei, D. M. Probabilistic topic models. *Commun. ACM* 55 (4): 77–84, 2012.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *Journal of machine Learning research* 3 (Jan): 993–1022, 2003.

Chan, K. R., Lou, X., Karaletsos, T., Crosbie, C., Gardos, S., Artz, D., and Rätsch, G. An empirical analysis of topic modeling for mining cancer clinical notes. In *13th IEEE ICDMW*. IEEE, pp. 56–63, 2013.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., and Blei, D. M. Reading tea leaves: How humans interpret topic models. In *Proceedings of the 23th NIPS*. pp. 288–296, 2009.

Charmaz, K. *A construção da teoria fundamentada: guia prático para análise qualitativa*. Bookman Editora, 2009.

Chen, J., Wei, W., Guo, C., Tang, L., and Sun, L. Textual analysis and visualization of research trends in data mining for electronic health records. *Health Policy and Technology* 6 (4): 389–400, 2017.

Chen, Y., Bordes, J.-B., and Filliat, D. An experimental comparison between nmf and lda for active cross-situational object-word learning. In *2016 Joint IEEE ICDL-EpiRob*. IEEE, pp. 217–222, 2016.

Chen, Y., Ghosh, J., Bejan, C. A., Gunter, C. A., Gupta, S., Kho, A., Liebovitz, D., Sun, J., Denny, J., and Malin, B. Building bridges across electronic health record systems through inferred phenotypic topics. *Journal of Biomedical Informatics* vol. 55, pp. 82 – 93, 2015.

Chertow, G., Soroko, S., Paganini, E., Cho, K., Himmelfarb, J., Ikizler, T., and Mehta, R. Mortality after acute renal failure: Models for prognostic stratification and risk adjustment. *Kidney International* 70 (6): 1120–1126, 2006.

da Silva, A., Hummel, J. R., Cabral, T. S., Carvalho, C. C. R., and Busanello, J. Índices de sedação e ventilação mecânica em paciente sob cuidados intensivos. In *Salão Internacional de Ensino, Pesquisa e Extensão*. Vol. 11. Unipampa, 2020.

Dare, A. J., Fu, S. H., Patra, J., Rodriguez, P. S., Thakur, J. S., and Jha, P. Renal failure deaths and their risk factors in india 2001–13: nationally representative estimates from the million death study. *The Lancet Global Health* 5 (1): 89–95, 2017.

Denaxas, S. C., Asselbergs, F. W., and Moore, J. H. The tip of the iceberg: challenges of accessing hospital electronic health record data for biological data mining. *BioData Mining* 9 (29), 2016.

Ding Cheng, L., Thermeau, T., Chute, C., and Liu, H. Discovering associations among diagnosis groups using topic modeling. In *AMIA Joint Summits on Translational Science*. pp. 43–49, 2014.

Doshi-Velez, F., Ge, Y., and Kohane, I. Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics* 133 (1): e54–e63, 2014.

Duarte, D. and Ståhl, N. Machine learning: a concise overview. In *Data Science in Practice*, A. Said and V. Torra (Eds.). Springer, pp. 27–58, 2019.

Fung, B. C., Wang, K., and Ester, M. Hierarchical document clustering using frequent itemsets. In *Proceedings of the 2003 SIAM international conference on data mining*. SIAM, pp. 59–70, 2003.

Gotz, D., Sun, J., Cao, N., and Ebadollahi, S. Visual cluster analysis in support of clinical decision intelligence. In *AMIA Annual Symposium Proceedings*. Vol. 2011. American Medical Informatics Association, pp. 481, 2011.

Guariguata, L., Whiting, D. R., Hambleton, I., Beagley, J., Linnenkamp, U., and Shaw, J. E. Global estimates of diabetes prevalence for 2013 and projections for 2035. *Diabetes research and clinical practice* 103 (2): 137–149, 2014.

Huang, S., Niu, Z., and Shi, Y. Product features categorization using constrained spectral clustering. In *International Conference on Application of Natural Language to Information Systems*. Springer, pp. 285–290, 2013.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., and Zhao, L. Latent dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications* 78 (11): 15169–15211, 2019.

Johnson, A. E., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. MIMIC-III, a freely accessible critical care database. *Scientific data* vol. 3, pp. 1–9, 2016.

Kalankesh, L., Weatherall, J., Ba-Dhfari, T., Buchan, I. E., and Brass, A. Taming ehr data: using semantic similarity to reduce dimensionality. In *MedInfo*. pp. 52–56, 2013.

Kane, R. L., Shamliyan, T. A., Mueller, C., Duval, S., and Wilt, T. J. The association of registered nurse staffing levels and patient outcomes: systematic review and meta-analysis. *Medical care* 45 (12): 1195–1204, 2007.

Kim, S., Kim, W., and Park, R. W. A comparison of intensive care unit mortality prediction models through the use of data mining techniques. *Healthcare informatics research* 17 (4): 232–243, 2011.

Koye, D. N., Magliano, D. J., Nelson, R. G., and Pavkov, M. E. The global epidemiology of diabetes and kidney disease. *Advances in Chronic Kidney Disease* 25 (2): 121 – 132, 2018. Diabetic Kidney Disease (c. 2018).

Kuang, D., Choo, J., and Park, H. Nonnegative matrix factorization for interactive topic modeling and document clustering. In *Partitional Clustering Algorithms*. Springer, pp. 215–243, 2015.

Lau, J. H., Newman, D., and Baldwin, T. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th EACL*. pp. 530–539, 2014.

Lehman, L.-w., Long, W., Saeed, M., and Mark, R. Latent topic discovery of clinical concepts from hospital discharge summaries of a heterogeneous patient cohort. In *36th Annual International Conference of the IEEE EMBS*. IEEE, pp. 1773–1776, 2014.

Lehman, L.-w., Saeed, M., Long, W., Lee, J., and Mark, R. Risk stratification of ICU patients using topic models inferred from unstructured progress notes. In *AMIA annual symposium proceedings*. Vol. 2012. American Medical Informatics Association, pp. 505, 2012.

Lu, H.-M., Wei, C.-P., and Hsiao, F.-Y. Modeling healthcare data using multiple-channel latent dirichlet allocation. *Journal of biomedical informatics* vol. 60, pp. 210–223, 2016.

Luo, M., Nie, F., Chang, X., Yang, Y., Hauptmann, A., and Zheng, Q. Probabilistic non-negative matrix factorization and its robust extensions for topic modeling. In *31st AAAI conference on artificial intelligence*, 2017.

Meskó, B., Drobni, Z., Bényei, É., Gergely, B., and Győrffy, Z. Digital health is a cultural transformation of traditional healthcare. *Mhealth* vol. 3, pp. 3–38, 2017.

Mihaela Coroiu, A., Delia Călin, A., and Nuţu, M. Topic modeling in medical data analysis. Case study based on medical records analysis. In *2019 International SoftCOM*, 2019.

M'sik, B. and Casablanca, B. M. Topic modeling coherence: A comparative study between lda and nmf models using covid'19 corpus. *International Journal* 9 (4), 2020.

Perotte, A. J., Wood, F., Elhadad, N., and Bartlett, N. Hierarchically supervised latent dirichlet allocation. In *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger (Eds.). Curran Associates, Inc., pp. 2609–2617, 2011.

Piskorski, J. and Yangarber, R. Information extraction: Past, present and future. In *Multi-source, Multilingual Information Extraction and Summarization*, T. Poibeau, H. Saggion, J. Piskorski, and R. Yangarber (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 23–49, 2013.

Řehůřek, R. and Sojka, P. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, pp. 45–50, 2010.

Röder, M., Both, A., and Hinneburg, A. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. WSDM '15, 2015a.

Röder, M., Both, A., and Hinneburg, A. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. Association for Computing Machinery, USA, pp. 399–408, 2015b.

Roque, F. S., Jensen, P. B., Schmock, H., Dalgaard, M., Andreatta, M., Hansen, T., Søeby, K., Bredkjær, S., Juul, A., Werge, T., et al. Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Comput Biol* 7 (8): e1002141, 2011.

Sena, M. R. D., Chahini, M., Braum, M. K., de Lima, S. M. M., Pimentel, S. K. S., Siqueira, V. A., et al. Mortalidade neonatal em hospitais públicos de alta e média complexidade no baixo amazonas. *Revista Eletrônica Acervo Saúde* 12 (5): e2286, 2020.

Steyvers, M. and Griffiths, T. Probabilistic topic models. In *Handbook of latent semantic analysis*, T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch (Eds.). Laurence Erlbaum Associates, 21, pp. 424–440, 2007.

Suri, P. and Roy, N. R. Comparison between LDA & NMF for event-detection from large text stream data. In *2017 3rd CICT*. IEEE, pp. 1–5, 2017.

Valenti, A. P., Chita-Tegmark, M., Tickle-Degnen, L., Bock, A. W., and Scheutz, M. J. Using topic modeling to infer the emotional state of people living with parkinson's disease. *Assistive Technology*, 2019.

Xie, P. and Xing, E. P. Integrating document clustering and topic modeling. In *Proceedings of the Twenty-Ninth Conference Uncertainty In Artificial Intelligence*. Association for Uncertainty in Artificial Intelligence (AUAI), 2013.

Yadav, P., Steinbach, M., Kumar, V., and Simon, G. Mining electronic health records (EHRs) a survey. *ACM Computing Surveys (CSUR)* 50 (6): 1–40, 2018.

Zhang, Y., Jiang, R., and Petzold, L. Survival topic models for predicting outcomes for trauma patients. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*. IEEE, pp. 1497–1504, 2017.

Zhao, J., Feng, Q., Wu, P., Warner, J. L., Denny, J. C., and Wei, W.-Q. Using topic modeling via non-negative matrix factorization to identify relationships between genetic variants and disease phenotypes: A case study of lipoprotein(a) (LPA). *PLOS ONE* vol. 14, pp. 1–15, 02, 2019.

APPENDIX A. DISCOVERED TOPICS

Table VII: Topics 1-6 from death collection.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
0.017*"skin"	0.031*"valve"	0.020*"liver"	0.009*"arrest"	0.014*"respiratory"	0.030*"hemorrhage
0.013*"drain"	0.026*"aortica"	0.016*"bleed"	0.008*"transfer"	0.011*"ventilation"	0.025*"head"
0.012*"wound"	0.021*"ventricular"	0.015*"renal"	0.007*"daily"	0.008*"wean"	0.011*"contrast"
0.011*"draining"	0.017*"mitral"	0.011*"cirrhosi"	0.007*"comfort"	0.008*"neuro"	0.011*"neuro"
0.011*"fractur"	0.014*"leaflet"	0.009*"hepatic"	0.007*"pulse"	0.008*"secretion"	0.010*"intubated"
0.009*"open"	0.013*"systole"	0.008*"ascites"	0.006*"arrive"	0.007*"intubated"	0.010*"frontal"
0.009*"ventilation"	0.012*"wall"	0.008*"sepsis"	0.006*"rhythm"	0.007*"thick"	0.010*"seizure"
0.008*"thick"	0.011*"mild"	0.008*"dialysis"	0.006*"unresponse"	0.007*"suction"	0.009*"cerebral"
0.008*"respiratory"	0.010*"regurgit"	0.008*"lactulos"	0.006*"received"	0.007*"shift"	0.009*"subarachnoid"
0.008*"suction"	0.010*"mildline"	0.007*"transplant"	0.006*"known"	0.006*"urin"	0.009*"ventricles"

Table VIII: Topics 7-11 from death collection.

Topic 7	Topic 8	Topic 9	Topic 10	Topic 11
0.012*"cmho"	0.012*"renal"	0.025*"contrast"	0.014*"pleural"	0.010*"lasix"
0.011*"renal"	0.009*"chronic"	0.013*"abdomen"	0.013*"pneumonia"	0.010*"chronic"
0.011*"ventilation"	0.009*"abdomin"	0.012*"catheter"	0.010*"unchange"	0.009*"renal"
0.010*"intubated"	0.008*"drain"	0.012*"liver"	0.010*"opacities"	0.007*"urin"
0.009*"sedated"	0.007*"rhythm"	0.011*"identifier"	0.010*"interval"	0.007*"hypotension"
0.008*"peep"	0.007*"breath"	0.010*"within"	0.009*"lobe"	0.007*"bacteremia"
0.007*"shock"	0.007*"heparin"	0.010*"vein"	0.008*"pneumothorax"	0.007*"pulse"
0.006*"balance"	0.007*"afib"	0.009*"abdomin"	0.007*"upper"	0.007*"infection"
0.006*"pulse"	0.006*"ventilation"	0.009*"pelvis"	0.007*"comparison"	0.007*"transfer"
0.006*"breath"	0.006*"neurologic"	0.008*"evidence"	0.006*"worsen"	0.006*"skin"

Table IX: Topics 1-6 from discharge collection.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
0.030*"valve"	0.012*"insulin"	0.016*"sepsis"	0.017*"respiratory"	0.009*"acute"	0.022*"head"
0.024*"aortica"	0.010*"surgeri"	0.013*"baby"	0.011*"ventilation"	0.008*"urin"	0.022*"fractur"
0.019*"arteria"	0.009*"wound"	0.013*"murmur"	0.010*"secretions"	0.008*"pulse"	0.019*"hemorrhage"
0.018*"ventricular"	0.008*"post-op"	0.013*"nicu"	0.009*"wean"	0.007*"bleed"	0.018*"contrast"
0.016*"mitral"	0.007*"regular"	0.013*"newborn"	0.009*"neuro"	0.007*"respiratory"	0.010*"neuro"
0.013*"leaflet"	0.007*"dilaudid"	0.013*"feed"	0.009*"thick"	0.007*"fluid"	0.010*"arteria"
0.012*"coronary"	0.007*"extreme"	0.011*"active"	0.008*"skin"	0.007*"rhythm"	0.010*"radiolog"
0.011*"cardiac"	0.007*"incision"	0.011*"born"	0.008*"urin"	0.007*"renal"	0.008*"evidence"
0.011*"systole"	0.006*"diet"	0.010*"week"	0.008*"intubated"	0.007*"chronic"	0.008*"mass"
0.010*"wall"	0.006*"intact"	0.010*"screen"	0.007*"suction"	0.006*"stool"	0.008*"hematoma"

Table X: Topics 7-11 from discharge collection.

Topic 7	Topic 8	Topic 9	Topic 10	Topic 11
0.020*"effusion"	0.017*"liver"	0.034*"tube"	0.054*"feed"	0.015*"respiratory"
0.016*"pleural"	0.015*"contrast"	0.013*"drain"	0.022*"active"	0.014*"tube"
0.013*"tube"	0.012*"fluid"	0.011*"draining"	0.022*"stool"	0.011*"acute"
0.012*"pulmonary"	0.011*"abdomen"	0.011*"pleural"	0.015*"respiratory"	0.010*"ventilation"
0.011*"radiolog"	0.011*"abdominal"	0.010*"pneumothorax"	0.014*"murmur"	0.010*"fluid"
0.011*"pneumonia"	0.010*"renal"	0.010*"effusion"	0.011*"retract"	0.009*"intubated"
0.009*"lobe"	0.010*"vein"	0.010*"wean"	0.011*"cpap"	0.009*"failure"
0.009*"interval"	0.009*"bleed"	0.009*"radiolog"	0.010*"benign"	0.008*"balance"
0.009*"lower"	0.009*"hepatic"	0.009*"respiratory"	0.010*"neonatolog"	0.008*"breath"
0.008*"opacities"	0.009*"radiolog"	0.007*"neuro"	0.010*"week"	0.007*"nutritional"