ABSTRACT

RUS, ANNISA MARLIN MASBAR. A Comprehensive Exploration of Topic Modeling Applications to Inform Decision-Making in Healthcare. (Under the direction of Dr. Maria Mayorga).

The use of Natural Language Processing (NLP) in healthcare, especially in the analysis of medical records, has great potential for improving patient outcomes and diagnostic models. However, accessing clinical notes from Electronic Health Records (EHR), which are often regulated and protected for patient privacy, poses challenges. In this dissertation, we explore alternative healthcare data resources beyond EHR data, focusing on internal hospital communications during the onset of the COVID-19 pandemic, call transcripts from a care coordination program for Diabetic Retinopathy (DR) patients, and journal papers related to the Colorectal Cancer Control Program (CRCCP). Our focus is on the use of an NLP method, topic modeling for different purposes.

In the first project, we analyze internal hospital communications during the early stages of the COVID-19 pandemic to identify the topics and their alignment with recommended communication topics from the American Hospital Association (AHA). Using structural topic modeling (STM), we analyzed 371 internal hospital communications, including emails and attachments from two health systems. Our study reveals nine topics, including visitor policies, best practice updates, and staff benefits. Additionally, we found that both hospital systems focused on operational adjustments. Lastly, our findings also indicate that the hospitals' communication contents were aligned with AHA recommendations.

In the second project, we explore factors influencing enrollment in a care coordination program for Diabetic Retinopathy (DR) patients. DR is a diabetes complication that can cause blindness. A care coordination and counseling program was offered to DR patients via telephone, but patient enrollment was low. Analyzing call transcripts from the enrollment process using STM, we identify topics related to program enrollment. To identify factors influencing patient enrollment, we developed three logistic regression models using the least absolute shrinkage and selection operator (lasso): one using call metadata variables (metadata), one with topics extracted from the call transcripts (topic-based), and one by adding metadata to the extracted topics (combination). The combination models achieved the highest performance, with the extracted topics enhancing predictive performance. The study found that discussions on scheduling counseling appointments have a positive association with enrollment, while conversations about program explanations have a negative association.

In the third project, we investigated the feasibility of a Supervised Extractive Summarization Method (SESM) for generating user-driven, multi-document summaries (QMDS). We used a

case study to assess its ability to capture trade-offs between service provision and promotion strategies in CRCCP programs. SESM leverages the efficiency of extractive summarization (ES) while incorporating the user-centric approach of content analysis (CA). To demonstrate the user-driven advantage, we compared SESM, which generates QMDS, to an Unsupervised Extractive Summarization Method (UESM) that creates generic summaries. The human evaluation indicated promise for SESM in generating summaries tailored to user needs. However, limitations remain. The current approach relies on pre-selected sentence pools, and automatic sentence labeling is needed for broader application. This study highlights the potential of extractive summarization, particularly the supervised approach (SESM), to streamline content analysis by generating focused summaries that target specific topics within the text. Future research directions include exploring multi-label classification for sentences and investigating the impact of different sentence embedding or topic modeling technique

Overall, this dissertation demonstrates the versatility of topic modeling, an NLP technique, for analyzing various healthcare data sources beyond Electronic Health Records (EHRs). Our studies explored internal hospital communications during a pandemic, call transcripts from a diabetic retinopathy program, and journal articles on colorectal cancer screening. The findings highlight the potential of NLP to uncover patterns, identify factors influencing patient behavior, and streamline content analysis for focused summaries. By overcoming challenges associated with accessing EHR data, this research paves the way for further exploration of NLP applications in healthcare data analysis to improve patient care and program effectiveness.

A Comprehensive Exploration of Topic Modeling Applications to Inform Decision-Making in Healthcare

by
Annisa Marlin Masbar Rus

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Industrial Engineering

Raleigh, North Carolina
2024

APPROVED BY:

Dr. Julie Ivy	Dr. Osman Özaltin
Dr. Min Chi	Dr. Maria Mayorga Chair of Advisory Committee

DEDICATION

To my parents, who taught me this:
"Berjuanglah sampai titik darah penghabisan."

BIOGRAPHY

Marlin was born and raised in Banda Aceh, Indonesia. Combining a passion for academia with a desire to improve healthcare, Marlin earned a Bachelor of Science in Industrial Engineering from Universitas Indonesia. Her interest in the intersection of technology and healthcare deepened during her Master of Science in Service Management and Design at the University of Warwick. This pursuit of knowledge culminated in her joining the esteemed graduate program at North Carolina State University in 2020. Marlin's research focuses on leveraging natural language processing (NLP) to advance healthcare practices.

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the unwavering support and guidance of many incredible people.

First and foremost, I am deeply grateful to my advisor, Dr. Maria Mayorga. Your mentorship has been instrumental in shaping me into the researcher I am today. Thank you for your patience, encouragement, and constant challenge to be a better and wiser academic. I am incredibly fortunate to have had you guide me through this challenging yet rewarding journey. I am also grateful for Dr. Julie Ivy, Dr. Osman Özaltin, and Dr. Min Chi for your feedback and guidance in this journey. I want to thank *Badan Pengembangan Sumber Daya* (BPSDM) Provinsi Aceh who has sponsored my PhD program until the end. This degree will not be possible without your support.

My deepest gratitude extends to my family. Dad (Prof. Dr.Raja Masbar) and Mom (Prof. Dr. Normalina Arpi), your unwavering love and encouragement have been my constant source of strength – especially during the low moments. Thank you for all your prayers and unwavering belief in me. To my siblings, Kak Mina, Bang Mubdi, Ina, and Adir, thank you for being my biggest cheerleaders and listening to my struggles. I am also grateful for the new members of the family, Bang Sangga, Kak Elle, Nurul, Yusuf, and Reia.

Finding a second family in Raleigh, North Carolina, has been a blessing. My heartfelt thanks go out to PERMIAS at NC State, including Mas Elvin, Mba Ines, Mas Yos, Mas Alberth, Mas Kiki, Iin, Arif, Mba Satya, Jizun, Bang Mindo, Mba Diesta, Mas Radit, Mas Azis, Mba Dilla, Novan, Anin, Mba Fit, Mas Eko, Mba Dwi, Mas Arif, Mba Tyas, Mas Panji, Mba Icha, Mas Aryo, Mba Rini, Mas Yudi, Troy, Indi, Prime, Arina, and Andrea. You weren't just friends who explored the US with me; you became my brothers and sisters during my time in Raleigh. Thank you for sharing your delicious Indonesian food and opening your homes.

My academic journey wouldn't have been complete without the support of incredible friends. Thank you, Whitney, Greg, Minerva, and Guillermo, for introducing me to American culture. To Erik, Margaret, Danika, Yiwei, Nasrin, Priscille, and Mitchell, your help with my studies and research has been invaluable.

Finally, to the love of my life, my husband Bang Evan Saputra. You came into my life during this PhD journey, and your presence has brought immense joy and laughter. Thank you for being my husband - a patient listener, a cheerleader, a culinary master who kept me well-fed, and a calming presence. You turned these busy, stressful years into an adventure filled with love, laughter, and memories I will cherish forever. I am incredibly grateful and fortunate to have you by my side. And for all my children to be, I dedicated this work for you.

Thank you all. I am very grateful for you all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER

1

INTRODUCTION

The use of natural language processing (NLP) in healthcare is increasing due to massive text data collection in healthcare. When NLP is applied to medical records, this tool can assist healthcare workers in predicting patient outcomes, enhancing hospital triage procedures, and creating diagnostic models for identifying early-stage chronic illnesses[63]. These uses could be especially valuable in critical care settings, where there is immense patient data to assess, and predicting patient mortality is a standard practice [63]. However, clinical notes are difficult to acquire because the documents are heavily regulated and protected from sharing patient data [36]. Publicly accessible data is available but limited. For instance, the Medical Information Mart for Intensive Care (MIMIC III) has documented electronic health records (EHR) from diverse U.S. health systems between 2001 and 2012, including clinical notes. According to a systematic review from Gao et al. [36], studies in 2021 relating to shared clinical notes mainly focused on Name Entity Recognition (NER) at the token-level and sentence classification, summarization, and information extraction at the document-level task. Therefore, there is value in exploring additional healthcare data sources for decision support or gaining insights.

1.1 Leveraging Text Data in Healthcare

Text data is abundantly available in healthcare and exists in various forms. Electronic Health Records (EHR) are the most commonly known text data that provide valuable information, and the popular NLP tasks that are conducted are information extraction and text categorization [63, 109]. To illustrate, unstructured data such as clinical notes can be utilized to predict non-violent patients in a psychiatry ward and understand the patient's experience of living with a particular disease [83, 44]. Moreover, integrating unstructured data with structured information (e.g., age, sex, medication group, ICD-10 code, etc.) holds potential for enhancing predictive capabilities in anticipating future incidents such as falls, severe maternal morbidity, and hospital readmissions [28, 20, 33, 38].

As previously mentioned, even though these clinical notes are proven to be essential data sources for the medical field, they are not always publicly available. Thus, the alternative is to leverage other text sources to gain insight from other stakeholders in healthcare systems. For instance, studies have used publicly available data such as 13,000 COVID-19 policies worldwide, newspaper articles, patient feedback, and social media in their research [41, 55, 31]. Based on these studies, the text data has proven to be a good alternative depending on the research goal.

In this dissertation, we aim to explore various healthcare data resources that can help us overcome the limitations of EHR data. Our analysis focuses on three different datasets: internal hospital communication records during the early stages of the COVID-19 outbreak, transcripts of calls during the enrollment process in a care coordination program for Diabetic Retinopathy (DR) patients, and scientific papers related to the Colorectal Cancer Control Program (CRCCP). Our primary objective is to use NLP techniques, specifically topic modeling, to gain insights from these datasets.

1.2 Overview of Remaining Chapters

In Chapter 2, we will discuss the topic modeling methods used throughout the remaining chapters, namely latent Dirichlet allocation (LDA) and structural topic modeling (STM). Both LDA and STM are unsupervised models and Bayesian inferential approaches that assume each document has a mixture of topics. Next, in Chapter 3, we will delve into the implementation of STM in extracting topics from internal hospital communications at the beginning of the COVID-19 pandemic. In Chapter 4, we shift our focus from extracting good topics from topic modeling to using the topics as a vector representation of documents for predicting outcomes. In this project, we aim to predict the probability of DR patients enrolling in a care coordination program. Lastly, in Chapter 5, we explored the adaptation of the topic modeling approach to generate a query-based summary from a multi-document composed of scientific papers. As a case study, we aim to summarize the trade-offs between provision and promotion strategies in CRCCP from the journal articles instead of creating a generic summary.

CHAPTER

2

TOPIC MODELING

In this section, we describe the general steps for topic modeling. Then, we discuss two Bayesian topic modeling techniques: latent Dirichlet allocation (LDA) and structural topic modeling (STM). LDA, introduced by Blei, Ng, and Jordan [11], is argued to be the most popular and simplest version of the Bayesian topics model [83, 85]. Later, Roberts, Stewart, and Airoldi [85] proposed STM as an extension of LDA, which takes into account the correlation between topics and document's metadata. We will first explain LDA and then move on to STM. Lastly, we explain the Gibbs Sampling approach which is used to draw samples.

2.1 General Steps for Topic Modeling

Topic modeling is a technique used to identify topics within a set of documents. It involves using statistical models to analyze unstructured text data. This field is rapidly evolving, with new methods being introduced frequently [83]. Topic modeling is a type of unsupervised natural language processing (NLP) model. Topic modeling results do not directly produce topic names. Instead, the top-n words with the highest probability per topic are used to represent each topic. The aim of topic modeling is to identify coherent topics with clear themes based on the top-n words in each topic. Generally, NLP will follow the steps shown in Table 2.1.

Table 2.1: Overall flow of the Topic Modeling Steps

Process	**Details**
Text preprocessing	The objective of this phase is to identify significant words that can be used for topic modeling. The text preprocessing includes: • Converting text to lowercase • Removing stopwords, such as I, am, and you. • Lemmatizing words or converting the words into root words, if necessary. • Tokenizing words. This process split the text into individual words. For example, text "health care news" will be tokenized as "health", "care", and "news". • Counting n-gram words. N-gram words are a contiguous sequence of n-words from a given sample of text or speech. For example, 2 grams (i.e. bigram) of "health care news" will be ("health", "care") and ("care","news"). • Combining standard terms that come together, such as "health" and "system" becomes "healthsystem". This process • Customize cleaning, depending on the case. This process could be removing names and other sensitive or unnecessary words from the text.
Topic Modeling	In this phase, we will extract topics from the documents using a topic model. Each document is assumed to have a mixture of topics, which will be referred as topic proportion. Before we begin the process, we need to determine the number of topics to extract that will be further explained in the next section.
Topic Interpretation	Topic model does not draw conclusions or interpretations. Typically, the top-n words with the highest probability of occurrence per topic are used to represent that particular topic. Subsequently, these word allow human to interpret these groups as topics based on a shared thematic association among the words. Consequently, the objective of topic interpretation is to assign a representative label to each word group. The topic interpretation can be conducted by analyzing the following: 1. The top-n words with the highest probabilities or exclusive score within each topic. In this dissertation, we use $n = 15$ or 20. 2. Document instances that strongly exemplify the topic. This could be identified by examining a document that has a high topic proportion of a particular topic. We will further discuss the topic proportion in Section 2.2.

2.1.1 Determining Number of Topics

Topic modeling requires the user to determine the number of topics, but there is still ongoing debate about the best way to do this. In this dissertation, we assessed topic performance based on semantic coherence and word exclusivity to identify high-quality topics as suggested by Roberts et al. [84].

Semantic coherence

Semantic coherence is maximized when the most likely words in each topic frequently co-occur [70]. For a pair of words v and v' in a topic, let $D(v, v')$ be the number of documents containing both words and $D(v)$ be the number of documents containing v. Then, for a list of M most probable words in topic k, the semantic coherence of topic k is calculated as:

$$C_k = \sum_{i,j \in \{1 \ldots M\},\ i \neq j} \log\left(\frac{D(v_i, v_j) + 1}{D(v_j)}\right) \tag{2.1}$$

We set M to 10 words. Note that logarithmic transformation can result in a negative value when the ratio in the bracket is less than one. A coherence score around zero suggests a higher coherence.

Exclusivity of Words

A word's exclusivity to a topic is assessed by its usage rate relative to a set of comparison topics (i.e., all topics or similar topics). In particular, let $\beta_{v,k}$ be the frequency of word v in topic k. Then, exclusivity of v to topic k relative to a set of comparison topics K is given by $\phi_{v,k} = \beta_{v,k} / \sum_{j \in K} \beta_{v,j}$. Since both frequency and exclusivity are important factors in determining a word's semantic content, we also consider a composite metric called FRequency and EXclusivity (FREX) proposed by Airoldi and Bischof [4].

FREX of a word v in topic k is the weighted harmonic mean of the word's exclusivity and frequency rank with respect to other words in the same topic. For a given weight ω for exclusivity, FREX is calculated by:

$$FREX_{v,k} = \left(\frac{\omega}{ECDF_{\phi}^{k}(\phi_{v,k})} + \frac{1-\omega}{ECDF_{\beta}^{k}(\beta_{v,k})}\right)^{-1} \tag{2.2}$$

where $ECDF_{\phi}^{k}$ and $ECDF_{\beta}^{k}$ are the empirical CDF of the words' exclusivity and frequency in topic k, respectively. We set $\omega = 0.7$ to favor exclusivity over frequency. The exclusivity of each topic is calculated by summing the exclusivity of the top 10 words in each topic.

2.2 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) was introduced by Blei et al. (2003) as a generative model for text that uses Bayesian inference to perform topic modeling. The main idea behind LDA is to represent documents as a mixture of topics and to model the generation of words in a document based on these topics. LDA aims to learn the underlying topic structure of a corpus by estimating two sets of parameters: (1) document-topic distribution for each document and (2) topic-word distribution for each topic. A document is generated by sampling a mixture of these topics (document-topics distribution) and then sampling words from that mixture(topic-word distribution). This model is not only applicable for text modeling but can also be broadly applicable to general collections of discrete data.

Consider

D	number of documents in the data set, index by $d \in 1,\ldots,D$
N_d	number of words in a document, index by $n \in 1,\ldots,N_d$
V	vocabulary of distinct terms, index by $v \in 1,\ldots, V$
K	number of topics, index by $k \in 1,\ldots,K$
β_k	probability mass function represents the frequency of words generated from a particular topic.

For a collection of documents D , each document d contains N_d words, V number of distinct terms, and K topics, indexed by k . Each topic k is associated with a V -dimensional probability mass function (pmf), β_k , that controls frequency according to which terms are generated from that topic. When a document d is generated, terms from V are assigned to each of the N_d positions. Instances of terms that fill these positions are typically referred to as the *words*. While the terms in V are unique, *words* in the generated document may have multiple occurrences of the same term.

Written out in full, the probability of a document is therefore expressed as:

$$p(w) = \int_{\theta_d} \left(\prod_{n=1}^{N_d} \sum_{k=1}^{K} p(w_{d,n}|\mathbf{B}z_{d,n}) p(z_{d,n}|\theta_d) \right) p(\theta_d) d\theta \tag{2.3}$$

where

$p(\theta_d)$	is K -dimensional Dirichlet distribution of document-topic distribution
$p(z_{d,n}|\theta_d)$	is K -dimensional multinomial distribution parameterized by θ_d
$p(w_{d,n}|\mathbf{B}z_{d,n})$	is V -dimensional multinomial distribution over the words.

The generative process of LDA is as follows:

1. For each document, draw a K -dimensional Dirichlet vector θ_d that capture the expected proportion of words in document d that can be attributed to each topic k. This K -dimensional Dirichlet vector θ_d is referred as the document-topic distribution.

2. For each *word* or position in the document, indexed by n , draw a topic for the associated word or position. We do this by sampling an indicator $z_{d,n}$ from a Multinomial$_K(\theta_d, 1)$.

3. For each word that has been associated with a topic, sample the actual words $w_{d,n}$ from a Multinomial$_V(\mathbf{B}\mathbf{z}_{d,n}, 1)$, where the matrix $\mathbf{B} = [\beta_1|...|\beta_K]$ encodes the distribution over terms in the vocabulary associated with the K topics.

2.3 Structural Topic Modeling (STM)

LDA by [11] is a well-known NLP approach used in healthcare. For example, several studies have applied LDA to extract topics from public opinion during the pandemic [31, 55, 59] and from clinical notes [38, 33, 83, 28]. As an unsupervised model, LDA extracts topics from a collection of documents or corpus. Each document is assumed to have a distribution of topics (document-topic distribution), and each topic is a distribution of words (topic-word distribution). For example, in a health article, the discussion of “disease symptoms" might be extensive, while “possible treatments" and “prevention of the disease" might be covered to a lesser extent, with the combined proportions summing up to one. These proportions are referred as topic proportion.

A significant limitation of the LDA is not considering topic correlations. For instance, a document discussing “vaccines” can be closely related to “visitor policy.” To address this issue, Blei and Lafferty [10] introduced correlated topic modeling (CTM). Later, Roberts, Stewart, and Airoldi [85] introduced STM, an extension of LDA and CTM, incorporating metadata such as author identification.

The incorporation of correlation between topics and metadata into the topic model are the main difference between LDA and STM. STM incorporate the correlation between topic by replacing Dirichlet distribution with logistic-normal distribution for the document-topic distribution. Moreover, STM incorporate the metadata covariate into the topic by model the mean vector of the logistic normal distribution as a simple linear mode such that $\mu_d = \Gamma' x'_d$.

In details, let

D	number of documents in the data set, index by $d \in 1,...,D$
N_d	number of words in a document, index by $n \in 1,...,N_d$
V	vocabulary of distinct terms, index by $v \in 1,..., V$
K	number of topics, index by $k \in 1,...,K$
β_k	probability mass function represents the frequency of words generated from a particular topic.
X	Matrix of covariates for topics in document-topic distribution with dimension $D * P$, where P is number of covariates. Each row of this matrix is denoted as $\mathbf{x}_d$.
Y	Matrix of covariates for words in topic-word distribution with dimension $D * A$, where A is number of covariates. Each row of this matrix is denoted as $\mathbf{y}_d$.

Following the same process as LDA, the STM is different in the following way

$\theta_d \sim \text{Logistic Normal}_{K-1}(\Gamma' x'_d, \Sigma)$,	is K -1 dimensional logistic normal distribution of document-topic distribution. We can represent the logistic normal by drawing $\eta_d \sim \text{Normal}_{K-1}(\mu_d, \Sigma)$ and mapping to the simplex by specifying $\theta_{d,k} = \exp(\eta_{d,k})/(\sum_{i=1}^{K} \exp(\eta_{d,i}))$, where $\mu_d = \Gamma' x'_d$
$\gamma_k \sim \text{Normal}_P(0, \sigma_k^2 I_P)$, for $k = 1...K-1$	is the mean of the logistic normal where $\Gamma = [\gamma_1 \vert ... \vert \gamma_K]$ is a $P x (K-1)$ matrix of coefficients fro the document-topic distribution
$z_{d,n} \sim \text{Multinomial}_K(\theta_d)$, for $n = 1...N_d$	is K dimensional multinomial distribution for drawing a topic for a word position
$w_{d,n} \sim \text{Multinomial}_K(B z_{d,n})$, for $n = 1...N_d$	is K dimensional multinomial distribution for drawing a term conditional on such the topic drawn before for a word position

The process of generating documents is as follows;

1. For each document, draw a K -1 dimensional logistic normal vector θ_d that captures the expected proportion of words in document d that can be attributed to each topic k . This K -dimensional Dirichlet vector θ_d is referred to as the document-topic distribution.

2. For each *word* or position in the document, indexed by n , draw a topic for the associated word or position. We do this by sampling an indicator $z_{d,n}$ from a $\text{Multinomial}_K(\theta_d, 1)$.

3. For each word that has been associated with a topic, sample the actual words $w_{d,n}$ from a $\text{Multinomial}_V(\mathbf{B}\mathbf{z}_{d,n}, 1)$, where the matrix $\mathbf{B} = [\beta_1|...|\beta_K]$ encodes the distribution over terms in the vocabulary associated with the K topics.

Noted that the covariate could also influence the topic-word distribution, but in this dissertation, we will only consider the covariate that influenced document-topic distribution.

2.4 Gibbs Sampling

Gibbs sampling is a Markov Chain Monte Carlo (MCMC) algorithm used for statistical inference, particularly in Bayesian statistics and machine learning [37]. It is named after the physicist J. Willard Gibbs. Gibbs sampling is employed to approximate complex probability distributions, especially when direct sampling is difficult. The key idea behind Gibbs sampling is that it simplifies the process of sampling from a joint distribution by breaking it down into a series of conditional distributions, which are often easier to handle.

Following is a brief overview of Gibbs sampling algorithm:

1. Setup: Consider a multivariate probability distribution that we want to sample from. Let's say the distribution involves multiple variables, and you're interested in sampling from the joint distribution of these variables.

2. Initialization: Start with an initial guess for the values of all the variables in the distribution.

3. Iterative Sampling: Iterate through each variable in the distribution and update its value based on the conditional distribution of that variable given the current values of all the other variables. In other words, each variable is sampled from its conditional distribution while keeping all other variables fixed.

4. Repeat: Repeat the process for a large number of iterations. After a sufficient number of iterations, the samples from the joint distribution converge to a stationary distribution.

5. Output: The samples obtained after the burn-in period (initial iterations that are discarded to allow the chain to reach the stationary distribution) can be used for inference and estimation of various statistics.

Gibbs sampling is widely used in Bayesian statistics for drawing samples from complex posterior distributions, especially when analytical methods are infeasible. It's a fundamental tool in the Bayesian statistical toolkit and has applications in fields such as machine learning, econometrics, and computational biology.

CHAPTER

3

COMPARATIVE ANALYSIS OF EMERGENCY RISK COMMUNICATIONS IN TWO US HEALTH SYSTEMS IN THE EARLY STAGES OF THE COVID-19 PANDEMIC USING TOPIC MODELING

3.1 Introduction

The WHO developed emergency risk communication (ERC) guidelines for countries facing disasters or disease outbreaks. These guidelines aim to empower individuals to make informed decisions to protect themselves, their families, and their communities [116]. Likewise, the Center for Disease Control and Prevention (CDC) introduced the crisis and emergency risk communication (CERC) framework to guide public health officials and communicators in effectively disseminating information during emergencies to protect public health. However, these guidelines do not specifically pertain to institutional or hospital-level communications, which are crucial in case of events like the COVID-19 pandemic. Communication in any outbreak of emerging disease is crucial because healthcare workers are at a heightened risk of infection during outbreaks of emerging diseases, given their close proximity to infected individuals [115]. In October 2020, the American Hospital Association (AHA) provided a general communication framework that hospital systems can adapt for internal usage [3]. They *recommended* nine topics for the COVID-19 pandemic management internal communications. However, to the best of our knowledge, there is no information on whether healthcare systems have effectively *implemented* these recommendations.

Limited research studies have investigated the content of hospitals' internal ERC during the COVID-19 pandemic. For instance, Doleman et al. [27] developed The Staff Perceptions of Pandemic Management Scale (SPPMS) as a survey instrument to measure staff perceptions of the effectiveness of managerial communications during COVID-19. The survey requested participants to list specific information they wished to have received, yet little is known about the content of the communications that were actually received. Kursite et al. [58] employed thematic analysis to extract five themes regarding internal communications among healthcare providers through semi-structured interviews. However, this method has the potential to overlook important topics during interviews and is dependent on participant recall.

In this study, we take a direct approach to examining the hospital-level internal ERCs. We investigate internal communication data from two hospital systems. The first hospital system serves seven counties across three states and is located in an urban area, while the second serves three rural counties in one state. Davoodi, Healy, and Goldberg [23] found a significant difference between rural and urban hospitals' medical surge capacities. Another study reported that rural hospitals experienced a more significant ICU bed shortage than urban hospitals [106]. We employ structural topic modeling (STM), a natural language processing (NLP) method. The STM incorporates the communication source (i.e., hospital system) as the metadata, which enables us to compare the content of hospital-level ERC topics between a rural and an urban hospital system.

3.2 Methods

This section describes the data, briefly explains the STM and how we translate the results of the STM into topics in a systematic way.

3.2.1 Data collection

We analyze internal communications from the Coronavirus Task Force (CT) of two hospital systems. These communications are hospital-level ERCs sent to all staff via email. They cover several subjects, such as the changes in operational protocols, the latest treatment guidelines, and staff resources. Hospital System 1 (HS1) is an academically affiliated ten-hospital network with over 300 outpatient clinics serving seven counties across three states and located in urban areas. Hospital System 2 (HS2) is a network of five hospitals with 80 outpatient clinics across three rural counties in one state. The internal communications span from March 15, 2020 to March 27, 2021 for HS1 and from February 16, 2020 to May 22, 2021 for HS2. HS1 shared 267 communications, and HS2 shared 105 communications with us. All communications are in text form. The median document length is 276 words with an interquartile range between 154 to 462 words. This study was conducted with the approval of each hospital system's institutional review board.

HS1 has three types of communication: 1) clinical documents, 2) email text, and 3) email attachments, while HS2 only shared the email content in the form of Word documents. Clinical documents are disseminated exclusively to employees through local distribution channels and archived on the organization's internal web page, distinct from email correspondence (see Figure 3.1). We used Optical Character Recognition (OCR) from the "pytesseract" Python package to extract the text from these documents as input for the STM [49].

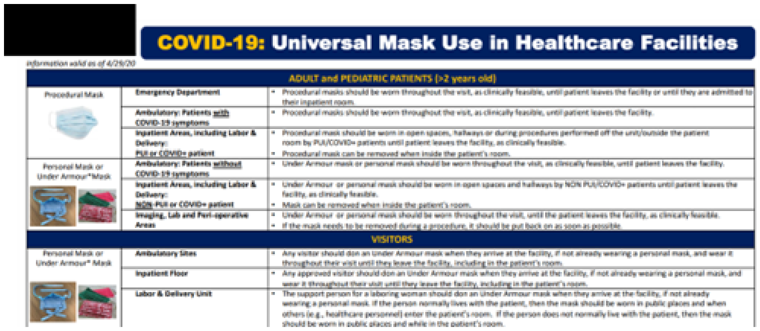
3.2.2 Text preprocessing

The unstructured text extracted from the documents was processed to prepare the input for STM. This process was time-consuming, but it was necessary to reduce the corpus size while keeping the words that reflected the main content of the documents. First, we converted all text into lowercase, split the text into tokens of individual words (tokenize), and created a bag of words for each document. Second, we removed the stop words, such as determiners (e.g., a, an, the), punctuation, coordinating conjunctions (e.g., but, or), prepositions (e.g., at, on, in), and URLs. Lastly, we manually combined some words and replaced the hospitals' names with HS1 and HS2. These pre-processing steps utilized the "quanteda" R package [7].

3.2.3 Structural topic modeling (STM)

We utilized STM as specified in Section 2.3. We used HS1 and HS2 as metadata for comparing topics covered by the two health systems considered in this study.

We ran the STM algorithm by varying the topic number from three to fifteen with four

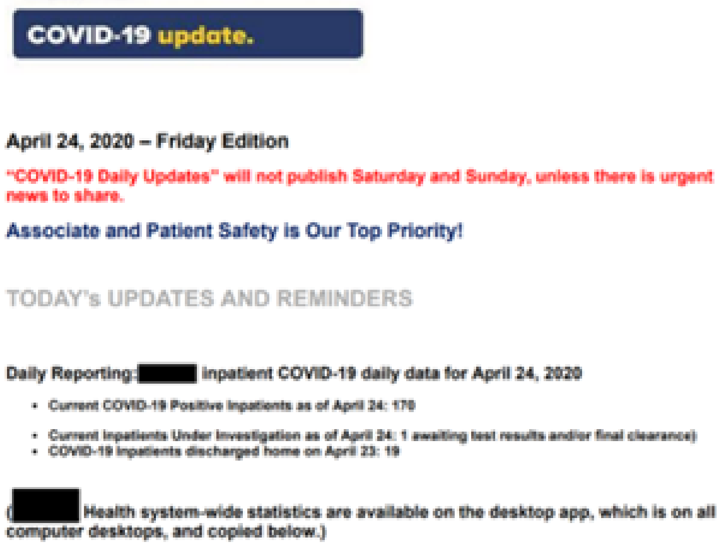


[redacted] **COVID-19: Universal Mask Use in Healthcare Facilities**

Information valid as of 4/29/20

ADULT and PEDIATRIC PATIENTS (>2 years old)		
Procedural Mask	**Emergency Department**	• Procedural masks should be worn throughout the visit, as clinically feasible, until patient leaves the facility or until they are admitted to their inpatient room.
	Ambulatory: Patients with COVID-19 symptoms	• Procedural masks should be worn throughout the visit, as clinically feasible, until patient leaves the facility.
	Inpatient Areas, including Labor & Delivery: PUI or COVID+ patient	• Procedural mask should be worn in open spaces, hallways or during procedures performed off the unit/outside the patient room by PUI/COVID+ patients until patient leaves the facility, as clinically feasible. • Procedural mask can be removed when inside the patient's room.
Personal Mask or Under Armour® Mask	**Ambulatory: Patients without COVID-19 symptoms**	• Under Armour mask or personal mask should be worn throughout the visit, as clinically feasible, until patient leaves the facility.
	Inpatient Areas, including Labor & Delivery: NON-PUI or COVID+ patient	• Under Armour or personal mask should be worn in open spaces and hallways by NON PUI/COVID+ patients until patient leaves the facility, as clinically feasible. • Mask can be removed when inside the patient's room.
	Imaging, Lab and Peri-operative Areas	• Under Armour or personal mask should be worn throughout the visit, until the patient leaves the facility, as clinically feasible. • If the mask needs to be removed during a procedure, it should be put back on as soon as possible.
VISITORS		
Personal Mask or Under Armour® Mask	**Ambulatory Sites**	• Any visitor should don an Under Armour mask when they arrive at the facility, if not already wearing a personal mask, and wear it throughout their visit until they leave the facility, including in the patient's room.
	Inpatient Floor	• Any approved visitor should don an Under Armour mask when they arrive at the facility, if not already wearing a personal mask, and wear it throughout their visit until they leave the facility, including in the patient's room.
	Labor & Delivery Unit	• The support person for a laboring woman should don an Under Armour mask when they arrive at the facility, if not already wearing a personal mask. If the person normally lives with the patient, then the mask should be worn in public places and when others (e.g., healthcare personnel) enter the patient's room. If the person does not normally live with the patient, then the mask should be worn in public places and while in the patient's room.

(a) Clinical document example from Hospital Systems 1



COVID-19 update.

April 24, 2020 – Friday Edition

"COVID-19 Daily Updates" will not publish Saturday and Sunday, unless there is urgent news to share.

Associate and Patient Safety is Our Top Priority!

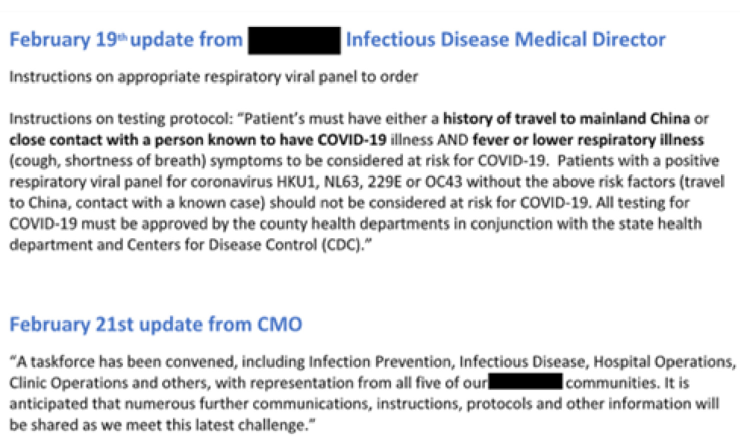
TODAY's UPDATES AND REMINDERS

Daily Reporting: [redacted] **inpatient COVID-19 daily data for April 24, 2020**

- **Current COVID-19 Positive Inpatients as of April 24: 170**
- **Current Inpatients Under Investigation as of April 24: 1 awaiting test results and/or final clearance)**
- **COVID-19 Inpatients discharged home on April 23: 19**

([redacted] **Health system-wide statistics are available on the desktop app, which is on all computer desktops, and copied below.)**

(b) Email example from Hospital Systems 1



February 19th update from [redacted] **Infectious Disease Medical Director**

Instructions on appropriate respiratory viral panel to order

Instructions on testing protocol: "Patient's must have either a **history of travel to mainland China** or **close contact with a person known to have COVID-19** illness AND **fever or lower respiratory illness** (cough, shortness of breath) symptoms to be considered at risk for COVID-19. Patients with a positive respiratory viral panel for coronavirus HKU1, NL63, 229E or OC43 without the above risk factors (travel to China, contact with a known case) should not be considered at risk for COVID-19. All testing for COVID-19 must be approved by the county health departments in conjunction with the state health department and Centers for Disease Control (CDC)."

February 21st update from CMO

"A taskforce has been convened, including Infection Prevention, Infectious Disease, Hospital Operations, Clinic Operations and others, with representation from all five of our [redacted] communities. It is anticipated that numerous further communications, instructions, protocols and other information will be shared as we meet this latest challenge."

(c) Email example from Hospital Systems 2

Figure 3.1: (a) Example of clinical documents that consist of tables and pictures, but in this study, we only analyze the text; (b) Example of email text from HS1; (c) Example of two emails from HS2

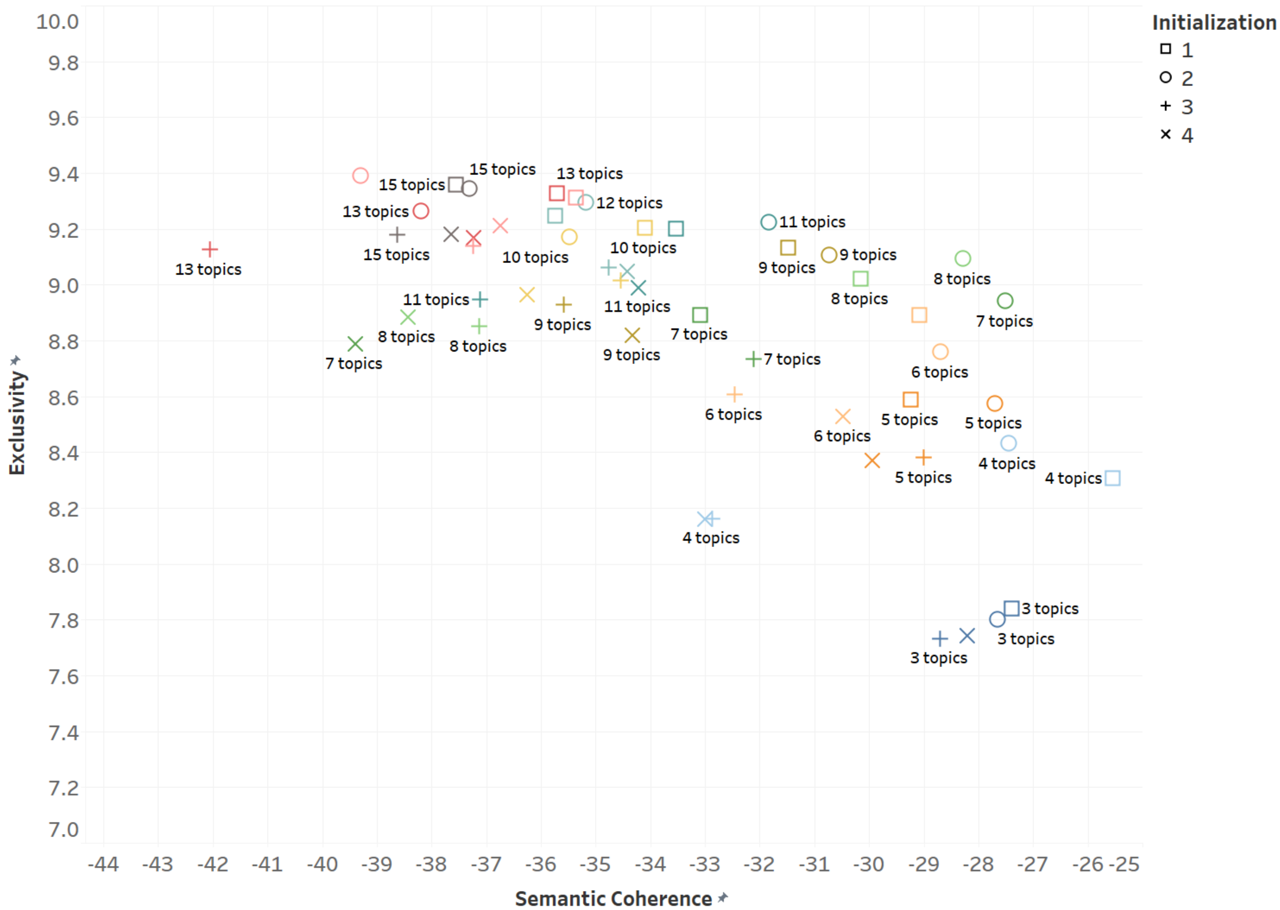


Figure 3.2: Semantic coherence and exclusivity value for three to fifteen topics generated by STM with four different initializations.

different initializations. Specifically, we initialized the STM model using LDA. Roberts, Stewart, and Airoldi [85] has demonstrated that utilizing LDA as an initial model, as opposed to a random initialization, facilitates faster convergence of the STM model. It is noteworthy that LDA's initialization process involves the random assignment of words, rendering each initialization unique due to the inherently multimodal nature of LDA.

We evaluated the performance of the number of topics based on semantic coherence and word exclusivity as specified in Section 2.1.1. We aim to identify topics with high semantic coherence and exclusivity. Figure 3.2 shows the centroid of exclusivity and coherence values of topics for different numbers of topics. Each initialization run is shown in a different shape. In Figure 3.2, values in the model with 6, 7, 8, and 9 topics are positioned in the upper right corner, which indicates that topics have high semantic coherence and exclusivity. Out of these models, we chose the 9-topics model because it is similar to the number of topics recommended by AHA.

3.2.4 Topic interpretation

We note that an STM model does not draw conclusions or interpretations. Instead, the resulting topics represent extracted word groups that frequently co-occur within documents. These groups are interpreted as topics based on a shared semantic association among the words. Topic interpretation aims to assign a representative label to each word group. We accomplish this step by analyzing the words with the highest probabilities within each topic and identifying document instances that strongly represent the topic. The representation of a topic within a document is determined by its topic proportions, wherein each document possesses proportions for all topics extracted from the model. For instance, in a model with three topics, a document will have a proportion associated with each of the three topics, and these topic proportions sum to one.

Table 3.1: Topic interpretation review process diagram.

Suggesting Topic Name	Seeking Consensus	Confirming with experts
Two main researchers met several times to assign names to 12 topics. For each topic, we examined : • The 20 words with high probability. • document examples with topics proportion more than 80% of the corresponding topics.	Three other researchers in the team reviewed the suggested topic names. • If consensus is not met, we pick a random example with a topic proportion of more than 50%. • Then, if consensus is still not reached, the topic will be discarded.	Two experts reviewed the consensus topics for confirmation.

We conducted a three-stage review process to increase confidence in topic interpretation, as depicted in Table 3.1. Initially, two researchers examined the words and proposed topic names. In particular, they reviewed the first 20 high-probability words and document examples with topic proportions of more than 0.8 of the corresponding topic. Then, three additional researchers reviewed these suggestions. If disagreements arose, the initial two researchers examined documents with at least 0.5 topic proportion and resolved the disagreements by revising topic names. We presented the revised topic names to two experts for confirmation.

These experts are employees within the considered health systems with relevant domain expertise. After the experts approved the topic names, further analysis was conducted.

3.3 Results

3.3.1 Nine topics of hospital-level emergency risk communications

We have identified nine hospital-level emergency risk communications topics across 371 emails and document attachments. Table 3.2 shows the topic names, the top 10 exclusive words, and the top 10 words with the highest probability in each topic, along with descriptions and snippet examples of example documents. Documents pertaining to a common topic generally originate from the same hospital system. In particular, topics 1,2,3,5,7,8 and 9 are mostly from HS1, topic 6 from HS2, and topic 4 is a mix of both hospital systems. It is important to note that there are more documents from HS1 than HS2.

Table 3.2: Identified nine topics of hospital-level emergency risk communications, FREX words, topic description, and snippet example from the documents.

Topic name	Shortened topic name	Description	Top 10 exclusive words (FREX order)	Top 10 words with high probability	Snippet Example
Vaccination	*Topic 1 Vaccination*	This topic focuses on early vaccine discussion	vaccine, vaccination, antibody, monoclonal, census, vaccinate, Pfizer, nonessential, self-screening, travel	vaccine, covid19, vaccination, HS1-state, travel, receive, work, schedule, provider, antibody	"... Please read through these important **COVID-19 vaccine** announcements and reminders so you can help keep staff, patients..." "... Emergency Use Authorization (EUA) for the use of the unapproved **monoclonal antibody** treatment Bamlanivimab for the treatment of mild to moderate..."

Table 3.2 (continued).

Topic name	Shortened topic name	Description	Top 10 exclusive words (FREX order)	Top 10 words with high probability	Snippet Example
Staff Appreciation	*Topic 2 Appreciation*	This topic pertains to staff appreciation	news, recognize, quality, heart, sender, delete, thank-you, privilege, spirit, orthopedics	team, email, thank-you, work, HS1-website, important, patient, help, service, day	"...Nationally recognized for clinical **quality** in **heart, orthopedics,** cancer and GI." "... **Thank you** for your understanding as we all work together through this time." "Their efforts in taking care of the whole person, the body, mind, **spirit** and soul, shows their dedication to the philosophy of nursing. I am very proud to **work** with such a great group of people.**Thank you** for all that you do"

Table 3.2 (continued).

Topic name	Shortened topic name	Description	Top 10 exclusive words (FREX order)	Top 10 words with high probability	Snippet Example
Patient safety procedure and resources for staff	*Topic 3 Patient Safety & Resources*	A combination of safety procedures for cleaning patient rooms and resources for staff, including food pantries and educational and training resources.	evs[a], designation, hydrogen-peroxide, Surfacide, live household[b], medium, approve, food pantry, clothes, FY20[c]	patient, room, use, care, HS1, PPE, eye-protection, approve, contact, need	"Bleach wipes will be **used** only for the cleaning of **rooms** of **patients** on enteric precautions and to clean specific equipment **used** for enteric patients. When you request bleach wipes from **EVS** for enteric precaution disinfection, you may receive a container that is different from the Clorox **hydrogen peroxide** wipes" "... Select your donation amount. Select "Other" in the **Designation** box. Type HS1-State Associate **Food Pantry**..."

Table 3.2 (continued).

Topic name	Shortened topic name	Description	Top 10 exclusive words (FREX order)	Top 10 words with high probability	Snippet Example
Educational and updated practice guidelines	*Topic 4 Education*	This topic covers updated guidelines on COVID-19 safety procedures for staff and updated practices among hospitals	symptom, guidance, COVID-19, exposure, precaution, restriction, outpatient, screen, face-masks, follow	COVID-19, patient, symptom, guidance, test, follow, mask, contact, hospital, day	"... **Test** results can be falsely negative if HCW (Healthcare worker) is asymptomatic. Exclude from work for 14 days unless **symptoms** develop during isolation, If so, **follow** instructions for symptomatic associate HCW, after any other **exposure** HCW may continue to work if asymptomatic. If symptomatic, refer to **guidance** for associates with **symptoms** consistent with **COVID-19** Self-monitor **symptoms**..."

Table 3.2 (continued).

Topic name	Shortened topic name	Description	Top 10 exclusive words (FREX order)	Top 10 words with high probability	Snippet Example
Staff resources and benefits	*Topic 5 Resources*	This topic covers ride-hailing services, COVID-19 clinic hours, and paid time off (PTO).	ride-hailing, COVID-19-clinic, balance, ride, vendor, PTO, protect, option, HS1-State, HEPA-machine[d]	HS1-state, available, ride-hailing, work, HS1, protect, information, use, PTO, department	"…Free discounted transportation for healthcare workers in **HS1-state** member hospitals. This [**ride-hailing**] voucher is good through Monday, June 1 …" "… Associates may use up to 120 hours of Advanced **PTO** (originally 80 hours) …" "… Never cover the**HEPA machine** with anything! **COVID-19** Clinic Hours. The **COVID-19 Clinic** …"

Table 3.2 (continued).

Topic name	Shortened topic name	Description	Top 10 exclusive words (FREX order)	Top 10 words with high probability	Snippet Example
PPE Guidelines	*Topic 6 PPE*	The topic content is about PPE management, guidelines, and updates. This include	HS2-internal website, facility, universal, turnaround, de-contamination, collection, conservation, list, committee, insider	policy, PPE, facility, update, care, use, mask, staff, work, employee	"… please see this Testing Guidance work instruction on the **HS2-internal website** page …" "… In addition, please remember that employees are not permitted to take **PPE** out of our **facility** for personal reasons.…" "..**UPDATE PPE** for COVID-19 Cohorted Patients When **caring** for confirmed COVID-19 patients sharing a semi-private room, …"

Table 3.2 (continued).

Topic name	Shortened topic name	Description	Top 10 exclusive words (FREX order)	Top 10 words with high probability	Snippet Example
Testing facilities	*Topic 7 Testing facilities*	Information about the testing facilities	tent, june, code, well-being, study, search, recover, program, portal, conference	test, see, occupational-health, june, call, important, wear, program, message, include	" ...clinician to nearest **OCCUPATIONAL HEALTH** office, on-site hospital clinic, or **testing tent** for **testing** ..." "... If you have further questions regarding how to appropriately **code** your staff labor, please review the document here or reach..." "..to be evaluated for symptoms and determine the need for COVID-19 **test**ing. The hours of the clinic are Monday-Friday 8 am to 4 pm. The clinic last day is **June** 12..."

Table 3.2 (continued).

Topic name	Shortened topic name	Description	Top 10 exclusive words (FREX order)	Top 10 words with high probability	Snippet Example
Ventilators	*Topic 8 Ventilator*	The topic content is clinical practice guidelines and sharing of practices for ventilators	intubation[e], non-PUI[f], ventilation, NICU[g], surgical, airway, nebulizer[h], disease, respiratory, child	patient, use, procedure, room, care, COVID-19, risk, COVID-19 positive, need, require	"…**Patients** who may have or are known to have covid19 who will require aerosol generating **procedures** like **intubation** or **nebulizer** therapy will continue to be **cared** for in airborne…" "…high **respiratory** tract pressurization and significantly longer duration of therapy than other category 1 or category 2 **procedures**. In the **NICU patient** population…"

Table 3.2 (continued).

Topic name	Shortened topic name	Description	Top 10 exclusive words (FREX order)	Top 10 words with high probability	Snippet Example
Visitor Policy	*Topic 9 Visitor*	This topic is regularly updated based on available knowledge at that time.	incident, bleach-wipe, commander, per-patient, enteric[i], garage, [name], exception, visitor, blue-jacket	visitor, hospital, COVID-19 positive, wear, food, support, provider, request, safety, time	"**...visitor**(s) to be approved Date(s) of approval length Reason for approval request COVID-19 status of the patient Chief Medical Officer [name], Chief Nursing Officer [name], and the **Incident Commander**s are the approvers and will reply to all with their decision..." "... **Hospital**-issued disposable head coverings, shoe coverings, and **blue jackets** are to be **worn** only in approved areas, ..." "...who have access to park in the main **hospital** parking **garage** may now temporarily enter through the front entrance of the..."

[a] EVS (Environmental Services)
[b] live-household (from live in the same household)
[c] FY20 (Fiscal year 2020, part of internal training name)
[d] HEPA-machine (High-efficiency particulate absorbing machine)
[e] intubation
[f] non-PUI (Non Person Under Investigation)
[g] NICU (Neonatal intensive care unit)
[h] nebulizer (drug delivery device)
[i] enteric

3.3.2 Topic proportions over time

We plotted and compared topic proportions with the percentage of total cases around the hospital location over time (Figures 3.3). In the figure, the first plot is for HS1, and the second is for HS2. The left vertical axis represents the average topic proportion of daily documents with bars, while the right vertical axis shows the percentage of cases with lines. The percentage of cases is calculated by dividing the number of cases reported on that day by the total cases. Total cases are limited to the duration of reported documents in the surrounding areas: seven counties across three states for HS1 and three counties in one state for HS2. Appendix 7 shows a similar figure for each topic. In this section, we present figures for topics occurring in both HS1 and HS2 (*Topic 4 Education*) to compare communication content between the hospitals.

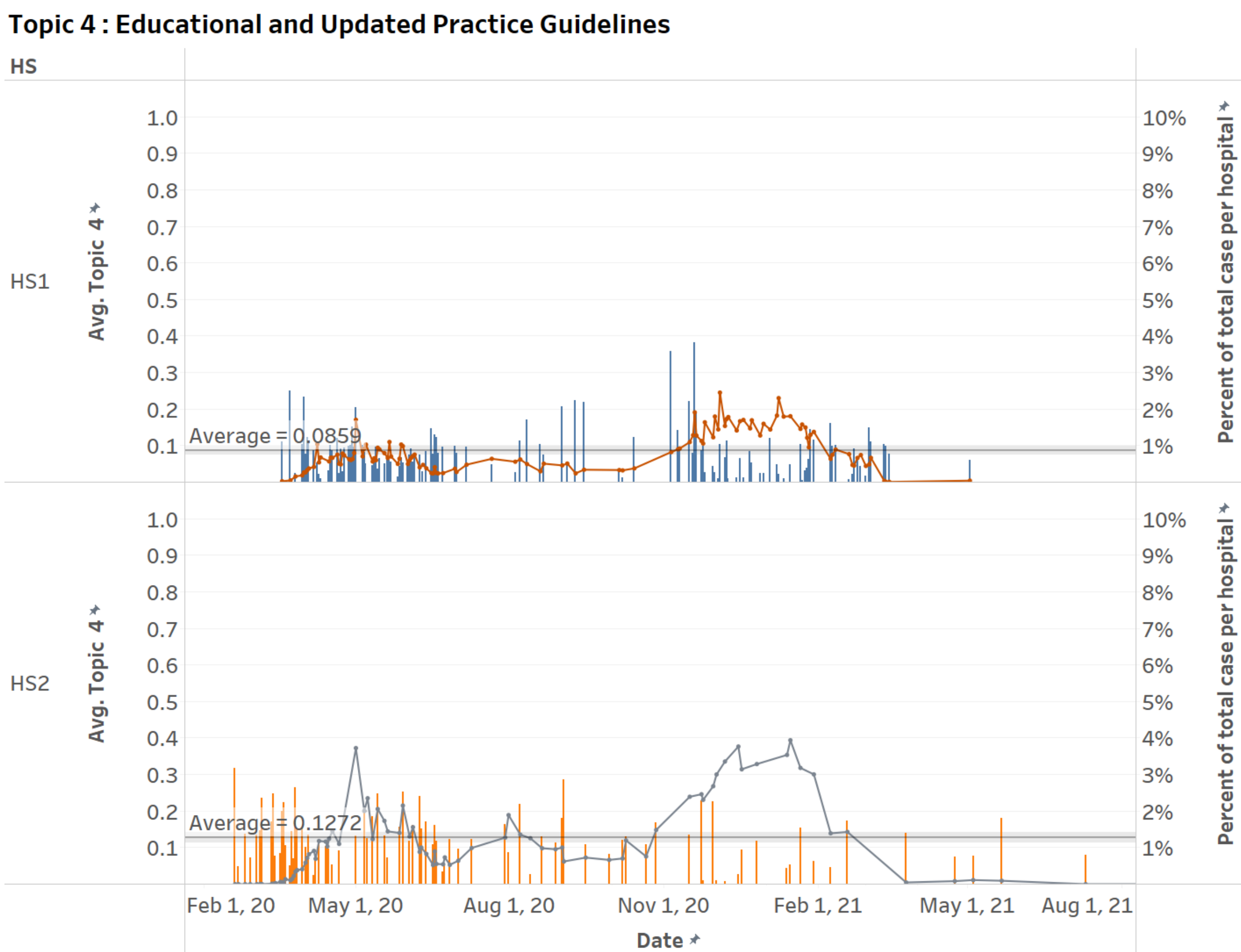


Figure 3.3: *Topic 4 Education* proportion and percent of total cases over time. The left vertical axis represents the average topic proportion of daily documents with bars, while the right vertical axis shows the percentage of cases with lines.

Both hospital systems discussed *Topic 4 Education* to some extent. HS1's average topic proportion across the time for this topic is 0.0859, while HS2's is 0.1272. However, HS2 started to cover this topic earlier and more intensively than HS1. Moreover, Figure 3.3 illustrates the decrease in *Topic 4 Education* discussion for HS2, which corresponds with the increase in COVID-19 cases between November 2020 to April 2021.

3.4 Discussion

Among the nine topics that we have identified, seven primarily originate from HS1, one from HS2, and one from both HS1 and HS2. In this section, we compare the topics generated by STM with topics recommended in AHA's general communication framework. We also discuss individual and shared topics, considering their content and timeline.

3.4.1 Comparison of topics generated by STM with those recommended by American Hospital Association

In October 2020, AHA provided a general messaging framework and communications toolkit that hospitals could adapt for internal and external use [3]. The AHA *recommended* nine pandemic-related topics to be communicated. In Figure 3.4, we align these topics with the ones that we identify using the STM model, which we considered as *implemented* communication topics. Seven of the AHA topics were included in the topics identified by the STM models.

The AHA topics do not directly match with STM topics due to the generality and specificity of the AHA topics. In regard to AHA generality, for instance, *Topic 1 Vaccination*, *Topic 4 Education*, and *Topic 8 Ventilators* in STM are merged into "clinical processes/ protocol" themes within AHA topics. Each of these STM topics discusses a specific clinical process related to vaccination, ventilators, and other treatments, while the AHA considers these as one general topic. Moreover, the AHA topic of "recognition, wellness, and resiliency" covers staff appreciation and resources, while the STM separates this topic into *Topic 2 Appreciation* and *Topic 5 Resources*. Furthermore, topics *Topic 6 PPE* and *Topic 7 Testing Facilities* correspond to AHA topics of "status of PPE supply" and "availability of tests" respectively, but AHA has a broader scope. For instance, topic *Topic 6 PPE* in STM discusses guidelines for using PPE, while AHA includes this discussion as a subtopic in "status of PPE supply". Likewise, *Topic 7 Testing Facilities* in STM focuses on providing information on testing facilities, while AHA covers both the availability of tests and its facilities in general.

On the other hand, several of the AHA topics are more specific than topics generated from the STM. For example, *Topic 5 Resources* covers different dimensions of healthcare workers,

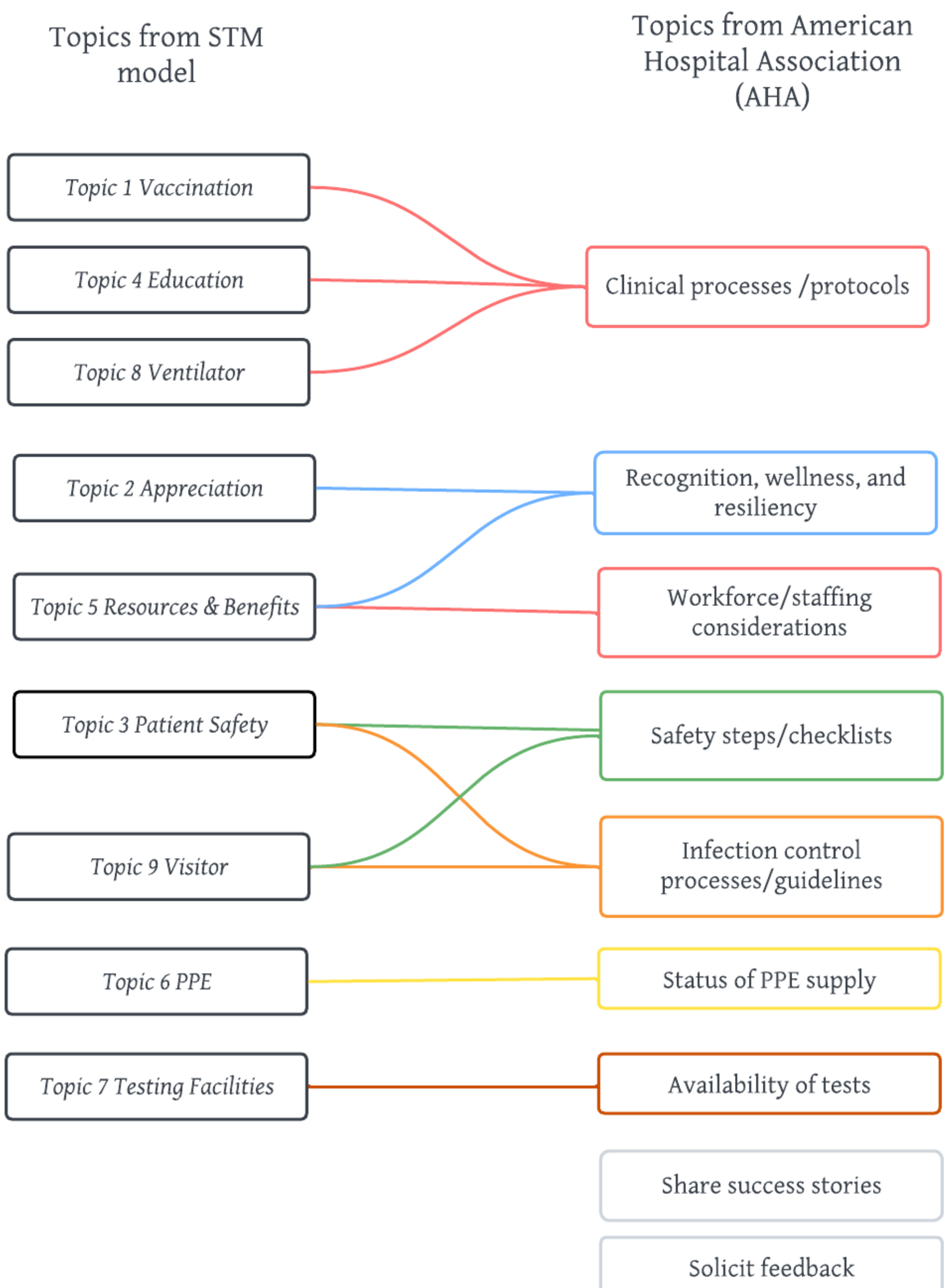


Figure 3.4: General message framework from American Hospital Association (AHA) and alignment with STM topics

resulting in two corresponding topics in AHA ("recognition, wellness, and resiliency" and "workforce/staffing considerations"). In the STM model, *Topic 5 Resources* encompasses staff resources (e.g., ride-hailing) and staff benefits (e.g., consideration for PTO). Moreover, *Topic 3 Patient Safety* corresponds to "safety steps/checklists" and "infection control processes" in the AHA. In STM, *Topic 3 Patient Safety* covers the proper use of PPE and safety procedures for environmental infection control measures. According to Collins (2008), PPE and environmental infection control measures are crucial methods for protecting patients. Similarly, *Topic 9 Visitor* also corresponds to "safety steps/checklists" and "infection control processes" in the AHA, but it focuses on visitors' perspective.

Interestingly, the AHA [3] particularly emphasized the significance of recognizing and acknowledging staff in multiple sections of their messaging framework. Likewise, the STM model also identifies this focus as a distinct topic in *Topic 2 Appreciation*. This alignment is aligned with broader research promoting staff appreciation during the pandemic. Smallwood et al. [94] argue that appreciation helps protect healthcare workers against moral distress, such as increased burnout, anxiety, depression, and post-traumatic stress disorder. Similarly, Honarmand et al. [50] found that 55% of respondents across 12 hospitals in Ontario, Canada, valued appreciation from leadership. Furthermore, lack of recognition of service employees who sanitized healthcare facilities as frontline responders was identified as one of the reasons for their elevated burnout risk [54].

As can be seen in Figure 3.4, "solicit feedback" and "share success stories" topics are not included in STM topics. However, these topics were not absent from the hospital communications. The STM model didn't identify the "solicit feedback" from healthcare staff topic because one of the health systems uses a separate channel for this purpose. Moreover, diverse "success stories" were shared in various different cases, making it challenging for the STM model to detect. For instance, one email showcased a 31-year-old patient's recovery following a 25-day hospitalization, while another email discussed the success of a pilot self-screening protocol. Overall, we believe that the content of the communications from the hospital systems that are captured by the STM aligns with the topics suggested by the AHA messaging framework.

3.4.2 Individual topics

Individual topics refer to the topics that are extensively discussed by specific health systems. Although one health system may give more importance to certain topics than others, it doesn't imply that the health system completely neglected the topic. Rather, it indicates that the health system discusses those topics to a lesser extent. The utilization of document sources as metadata in STM analysis has been observed to result in the prevalence of a particular health system's discussion on a given topic. We observed that HS1 frequently discussed topics related to COVID-19 treatment updates (*Topic 1 Vaccination* and *Topic 8 Ventilators*), staff appreciation (*Topic 2 Appreciation*), and operational adjustments (*Topic 3 Patient Safety*, *Topic 5 Resources*, *Topic 7 Testing Facilities*, and *Topic 9 Visitor*). On the other hand, HS2 focused primarily on *Topic 6 PPE*, which is part of the operational adjustment

In the early stages of the COVID-19 pandemic, it was crucial to provide reliable and accurate updates as the knowledge was rapidly evolving. Both health systems discussed this issue in *Topic 1 Vaccination* and *Topic 8 Ventilators*. Having the most up-to-date knowledge on COVID-19 was essential to control its spread, provide appropriate patient care, and safeguard healthcare workers who were in frequent contact with patients [105]. Moreover, Ow Yong et al. [77] believe that consistent, accurate, and concise updates are necessary to keep healthcare workers informed and motivated. In addition, by managing access to current research findings and accurate information, misinformation that may spread through social media could be mitigated [101]. Based on our result, HS1 and HS2 have discussed this topic group in a similar pattern over time, with HS1 having a higher average topic proportion than HS2. Notably, *Topic 1 Vaccination* gained significant attention towards the end of 2020, as vaccines became a more prominent discussion point. Conversely, the *Topic 8 Ventilators* was heavily discussed at the beginning of the pandemic, as it relates to treating patients with severe respiratory complications who require mechanical assistance to breathe using a ventilator.

In response to the continuous updates concerning the COVID-19 pandemic, hospitals have made operational adjustments to mitigate the virus transmission. HS1 and HS2 discussed the operational adjustment in *Topic 3 Patient Safety*, *Topic 5 Resources*, *Topic 6 PPE*, *Topic 7 Testing Facilities*, and *Topic 9 Visitor*. In the study conducted by McLean et al. [69], they found that focusing their adjustments on personal protective equipment (PPE), occupational health, and workplace safety was crucial in the emergency operation center of a large community hospital in Ontario, Canada. These adjustments were identified as key themes in operational changes. Our study further supports the occurrence of these themes in both health systems' communication content based on the topics extracted. Specifically, these topics related to operational adjustment to protect patients (covered in *Topic 3 Patient Safety* and *Topic 9 Visitor*) and healthcare workers (covered in *Topic 6 PPE* and *Topic 7 Testing Facilities*) from the virus. Noted that, these two practices are interdependent, as protocols created to protect patients also help protect healthcare workers and vice versa. In terms of scope, HS1 provides extensive coverage of *Topic 3 Patient Safety*, *Topic 7 Testing Facilities*, *Topic 6 PPE*, and *Topic 9 Visitor* over time, while HS2 primarily focuses on *Topic 6 PPE*. We observed that HS1 has more topics identified than HS2 due to the larger number of documents available from HS1 than from HS2. However, it is important to note that *Topic 6 PPE*, which mainly comes from HS2's document, encompasses the main measures to ensure the safety of both patients and healthcare workers in operational adjustment.

In addition, the management of human resources (i.e., *Topic 5 Resources*) during a crisis has been a widely discussed topic of great importance in operational adjustment. Various studies have emphasized this issue. For example, a survey conducted by Ow Yong et al. [77] found that neglecting staff welfare and human resources can have a negative impact on staff morale and perceived support during a crisis. In another study, Snyder et al. [96] found that staffing challenges, lack of facility support, and insufficient pay and incentives can cause burnout and turnover among certified nurses. Furthermore, McLean et al. [69] also identified human resources and workforce planning as crucial themes to be addressed during the hospital pandemic response. In our study, HS1 extensively discussed *Topic 5 Resources* during the pandemic, while HS2 covered a similar topic in a separate email chain that was not made available to us.

In regards to recognizing and appreciating the efforts of healthcare professionals, this matter is addressed in section 2. As we discussed earlier, implementing these practices can help protect healthcare workers from moral distress [94]. Conversely, the absence of such practices can result in job dissatisfaction, burnout, and high employee turnover rates among healthcare professionals [77, 96, 54].

In summary, the analysis of communication content from two health systems at the onset of the COVID-19 pandemic highlights the diverse range of topics discussed and the varying emphases placed on them. Both health systems addressed issues such as providing continuous COVID-19 updates, making operational adjustments, managing the workforce, and appreciating staff. Specifically, HS1 covers all topics to some extent, while HS2 focuses on PPE discussion. The difference in topic coverage can be attributed to the volume of available documents, rather than inherent priorities. These findings highlight the universal relevance of these topics across health systems and emphasize their importance in navigating public health crises effectively.

3.4.3 Shared topics

A shared topic refers to a topic that is discussed by both health systems with a marginal difference of less than 0.05 in the average topic proportion. We have identified one shared topic, *Topic 4 Education*.

In Figure 3.3, *Topic 4 Education* was discussed at the beginning of the pandemic, decreased in the middle, and re-surged from late November 2020 to March 2021. At the beginning of the pandemic, healthcare systems discussed the topic differently in the communication content because both health systems have different priorities. Documents from HS1 indicate that they first addressed the pandemic in early March by warning their personnel of false information circulating through news and social media. Tasnim, Hossain, and Mazumder [101] suggested that maintaining access to accurate and up-to-date research findings could help health systems tackle misinformation effectively. This strategy was implemented by HS1 by continuously updating the staff with credible COVID-19 updates. On the other hand, HS2 discussed the emergence of a novel virus from Wuhan and their first reported cases, then responded by developing a comprehensive plan as early as late February. Even though the first responses of hospital systems occurred at different times, both hospitals began sharing available practices immediately in their early communications.

In summary, this study has identified the majority of topics recommended by AHA. Additionally, both individual and shared topics that were identified have been widely discussed in previous research. Note that we confirmed the prevalence of these topics by directly analyzing the hospital's communication content using STM. In contrast, previous studies have relied on collecting surveys, analyzing focus group discussions, and conducting thematic analysis on emergency operation center documents to draw similar conclusions.

3.5 Conclusion

We investigated the internal emergency risk communications (ERCs) of two health systems during the COVID-19 pandemic. By analyzing the shared communications, we intended to mitigate biases associated with survey-based studies. Despite WHO's established ERC plans at the country level, ERC within hospitals, which is pivotal in patient care and staff well-being during the pandemic, remains under-explored.

We used structural topic modeling (STM), a Natural Language Processing (NLP) method, to extract nine topics from 371 email communications. After interpreting the topics through multi-level review stages, we found that seven topics were prevalent in Health System 1 (HS1), serving three states, while one was mainly discussed in Health System 2 (HS2), serving one state. Moreover, HS1 and HS2 discussed one topic concurrently.

Compared to the *recommended* topics from the AHA [3], our analysis using the STM model identifies the *implemented* communication topics by the health systems. Two topics recommended by AHA were not captured in our results either due to their coverage on other communication channels or because of variations in the presentation of their content. Furthermore, this study shows that the *implemented* communication contents of both health systems covered topics related to continuous COVID-19 updates, operational adjustments, workforce management, and staff recognition. There were variations in the topics covered and their proportion between HS1 and HS2. In particular, HS1 discussed all topics to some extent, while

HS2 focused more on discussing the guidelines for Personal Protective Equipment (PPE).

The present study provides insights into the potential of identified topics to inform health systems regarding their implementation of ERC approaches during pandemics. The study's outcomes can assist health systems in assessing their strategies, thereby contributing to improving the future management of pandemics. Additionally, this paper demonstrates NLP's ability to extract topics efficiently from documents.

This study is not without limitations. Results reflect data collected from a specific time frame and do not represent all communications used by the hospital systems. The recorded documents of the two health systems had different start and end points. Moreover, HS1 had more emails and attachments than HS2, which affected the topic extraction process. In the future, sentiment analysis can be used to evaluate ERCs throughout the pandemic. Furthermore, the topic proportions found in this study can be used as a feature to predict how different topics influence the ERC's positive and negative sentiment.

CHAPTER

4

PREDICTING PATIENT ENROLLMENT IN A TELEPHONE-BASED HEALTH CARE COORDINATION AND COUNSELING PROGRAM USING TOPIC MODELING

4.1 Introduction

Diabetic Retinopathy (DR) is a diabetes complication that can lead to vision-threatening issues (VTDR) and even blindness [14]. Detecting VTDR can be challenging due to its slow progression; thus, it is recommended that DR patients visit an eye doctor annually to ensure early detection [14]. Unfortunately, in 2020, only 58.3% of adults with diabetes received an eye exam, falling short of the 70.3% target set by Healthy People 2030 [75]. This concern has been labeled as "worsening" by the U.S. Department of Health and Human Services [75].

One favorable approach to help patients manage their eye care is through telephone-based health coaching (TBHC) and counseling. Previous research on diabetic patients indicates that TBHC is cost-effective and can assist patients with physical activity and dietary management but the implementation of these programs varies, such as different call frequencies or coaching content [42, 39, 88, 29, 45]. A US study involving 174,120 patients with various chronic conditions, including diabetes, also showed that generic TBHC with shared decision-making, patient information, self-management, and communication skills reduced hospitalizations, surgeries, and healthcare costs [114]. In a separate investigation, Wungrath and Autorn [117] observed that telephone-based counseling effectively enhanced medication adherence. In brief, TBHC and telephone-based counseling have shown promising advantages for patients.

One such telephone-based intervention program that focuses on assisting and counseling patients who are specifically at risk of DR is provided for patients by a care management company, and the program is called a care coordination and counseling (CC) program. This program, led by a certified ophthalmic assistant (COA), supports patients in managing their care through phone calls in between doctor visits. The COA could help patients with financial assistance programs, prescriptions and refills, insurance issues, transportation to visits, understanding their disease and care plan, and coordinating care with other providers. However, despite the service benefits, around half of the doctor-recommended patients did not enroll in the CC program. Therefore, the objective of this study is to explore factors associated with patients' decisions to either participate in or abstain from enrolling in the CC program.

Other researchers have explored factors associated with enrollment. For example, Haynes et al. [46]qualitatively explored a coaching program related to promoting physical activity for older people. Haynes et al. [46] explored the factors contributing to low participation by conducting a secondary thematic analysis of semi-structured interview data. However, this approach may be susceptible to overlooking certain interview topics and relies on participants' recall memory. Therefore, to mitigate these concerns, this study adopted a distinct approach by analyzing all of the conversational data (i.e., call transcripts) directly to extract relevant themes related to patients' enrollment in the CC program.

Given the substantial size of the data, employing natural language processing (NLP) for automated analysis of text could be beneficial in exploring the call transcripts. The CC company automatically transcribed the call interactions between patients and COA agents for program enrollment purposes. Reading these transcripts would be time-consuming. Thus, we employ NLP on these call transcripts to explore factors associated with not enrolling in the CC program. In this study, we propose to achieve this objective by transforming the call transcripts into vector representations which can serve as features for predicting program enrollment.

Call transcripts present challenges compared to other text documents due to their non-continuous nature, the inclusion of irrelevant “small talk," and grammatical errors [9]. However, these concerns are alleviated when employing topic modeling, such as Latent Dirichlet Allocation (LDA), which operates at the word level. Using topic modeling, we can transform the transcripts into vector representations and use them as features for prediction tasks. A few studies explored the use of call transcripts for prediction, but they lack interpretability because of black box approaches that are used, making it difficult to identify content or topics that are associated with the call outcome [119, 120]. In contrast, the topic modeling approach allows us to pinpoint the specific extracted topics that influence patient enrollment decisions.

Numerous healthcare studies have leveraged topic modeling to vectorize documents, such as clinical notes, and have paired the results with electronic health record (EHR) data, including demographic information, ICD-9 diagnosis codes, and laboratory tests, to facilitate prediction. Encouragingly, these studies have demonstrated promising performance [38, 33, 28]. Additionally, Rijcken et al. [83] have investigated various topic modeling approaches, such as LDA, specifically for prediction purposes. It is worth noting that previous studies have primarily concentrated on Latent Dirichlet Allocation (LDA). Structural Topic Modeling (STM) is a newer technique introduced by Roberts, Stewart, and Airoldi [85] that has received less attention in research compared to LDA. LDA does not consider the correlation between topics, whereas STM takes into account both the correlation between topics and document metadata, such as authorship when generating topics. Therefore, this study intends to investigate the atypical text format of call transcripts and explore the previously under-examined topic modeling method, STM, to produce document features for prediction.

In brief, the study objective is to explore factors associated with patient enrollment in a telephone-based counseling program, the CC program, from call transcripts collected during the enrollment process. To achieve this objective, we transformed enrollment call transcripts into vector representations by using STM to generate extracted topics that serve as document features. Subsequently, we predict enrollment by utilizing both these topics and call metadata and evaluated the predictive performance of three classification models.

4.2 Methods

This section will provide detailed information regarding the Care Coordination and Counseling (CC) program considered in this study, transcript data and the inclusion criteria, the chosen call metadata as predictive factors, an overview of topic modeling, and the configuration of the prediction model settings. Table 4.1 presents a visual representation of the overall research workflow. This study has been approved by the North Carolina State University institutional review board (IRB) under protocol # 12184.

Table 4.1: Overall flow of the methods used

Process	**Details**
Database preparation	The following are the data inclusion criteria: • Filter patients based on their presence in coaching calls after enrolling in the program. • Filter calls based on the transcripts' availability and call length.
Calls metadata preparation	Calls metadata that are used as predictors include: • **Total transcript length** in combined calls per patients • **Number of calls** received by patients in enrollment process. • **Interval** between the last call and the call before in days. • **Maximum of agent sentiment** • **Minimum of customer sentiment**
Text preprocessing	The objective of this phase is to identify significant words that can be used for topic modeling. The text preprocessing includes: • Converting text to lowercase • Removing stopwords, such as I, am, and you. • Lemmatizing words or converting the words into root words. • Tokenizing words.
Topic Modeling	In this phase, we applied Structural Topic Modeling (STM) to incorporate enrollment status in extracting topics from the call transcripts. • Metadata that is used is the enrollment status (1 = enroll, 0 = not enroll) • Number of topics analyzed is from 3 to 15 topics.
Classification models	Model details : • Logistic regression with Lasso is our classifier. • All independent variables are normalized. • Train and test data are split into 0.9 and 0.1, respectively. Models: • **Metadata model**: using call metadata as predictors. • **Topic-based model**: using topic proportions generated from STM as predictors. • **Topic + metadata model**: using topic proportions and call metadata as predictors.

Care coordination and counseling (CC) program

The CC program provides telephone-based counseling for patients with chronic eye conditions, aiming to bridge the gap between doctor's appointments and facilitating the technical aspects of eye care management. A care management company provides this service with a team that includes ophthalmology, biomedical engineering, data science, and healthcare services leaders. In contrast to telephone-based health coaching and counseling programs for diabetic patients, which provide general advice on physical activity and diet, this program specifically focuses on the patient's eye care management, including diabetic retinopathy, and assists patients with technical aspects (e.g., booking a doctor's appointments, coordinating with the pharmacy for medication, and facilitating transportation for going to the clinic).

The program provides services for financial assistance, prescriptions and refills, insurance issues, transportation to visits, understanding their disease and care plan, and coordinating care with other providers. Patients can have questions answered, concerns addressed, and even book doctor's appointments directly. These services are provided by certified ophthalmic assistants (COA) with an average of 12 years of in-clinic experience and specialized training in social work, case management, and motivational interviewing.

The CC program enrollment process begins with a doctor's appointment, during which the physician may recommend the program if deemed beneficial. Subsequently, a CC program agent contacts the doctor-recommended patients via phone to provide program details and extend the offer. Patients who consent to participate will receive future calls from the COA based on patients' availability. We refer to these interactions before the counseling begins as "enrollment calls", which are the focus of our study.

4.2.1 Data

Call inclusion criteria

For this study, we relied on transcripts of enrollment calls as our primary text data for topic modeling. These calls are recorded, transcribed automatically using Amazon Web Services (AWS), and reviewed by the care management agency to remove any personally identifiable information (PII). Furthermore, to ensure confidentiality, the dates of the calls were shifted for each patient, while maintaining the sequence and time between calls. It is worth noting that patients may have multiple calls before making a decision to enroll or decline participation in the CC program, as depicted in Fig 4.1. The recorded enrollment calls span from June 7th, 2021 to April 13th, 2023, and comprise around 30,662 calls, all in English.

We identify each call with a call ID, and its outcome is manually perceived and recorded by

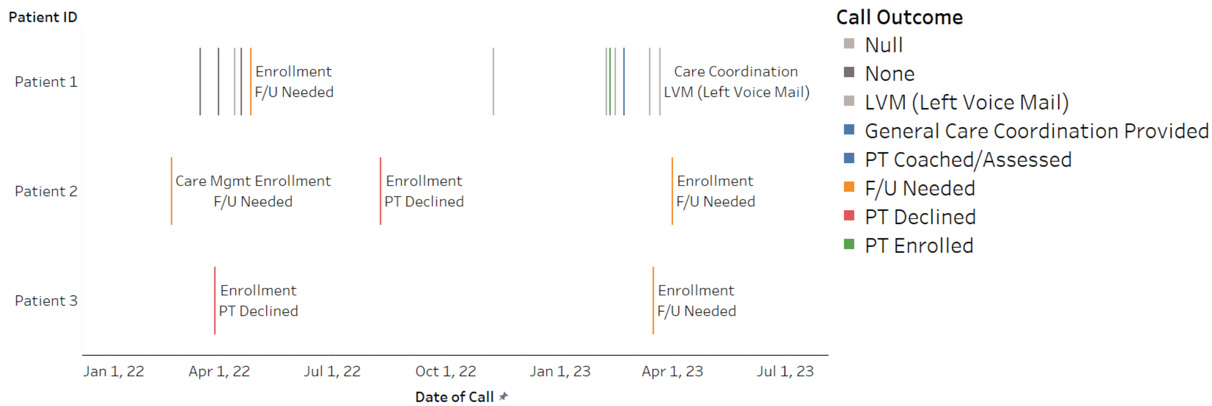


Figure 4.1: Call history example. Examples of three patients' call history over time. The y-axis shows the patient ID and the x-axis shows the date. Each patient has different colored bars that indicate different call outcomes. Each bar is labeled with the call type indicated on the first line of the label and the call outcome on the second line. However, not all call labelled because lines are too close to each other.

the program agent after the conversation, referred to as the "call outcome."For this study, we will only consider four distinct call outcomes related to enrollment calls. Other outcomes related to voicemail messages, transfers, calls disconnected, busy, leaving a message with someone, or invalid numbers will be excluded. However, we reserve the frequency of these calls to include in the total count of the number of calls. The four relevant outcomes are "Patient Enrolled" (PT Enrolled), "Patient Declined" (PT Declined), "Patient Unenrolled" (PT Unenrolled), and "Follow-Up Needed" (F/U Needed). For analytical simplicity, we have categorized these outcomes into two overarching groups: "enroll" (comprising PT Enrolled) and "not enroll" (comprising PT Declined, PT Unenrolled, and F/U Needed). Consequently, our dataset consists of a total of 9,382 call IDs meeting these specified criteria.

In this study, we further filter the text dataset (i.e., 9,382 call IDs) through a four-stage process.

1. **First stage: Participant selection**
   In the first stage, we applied a patient-level filter. Specifically, for patients who enrolled in the program, we focus on individuals who actively participated in the coaching program following enrollment, meaning that their enrollment was followed by a coaching call. This approach ensured that the included patients were deliberately interested in the program instead of merely expressing initial interest and subsequently disengaging from the agent. As for patients who did not enroll, we retained all their calls without exclusion.

2. **Second stage: Unavailable transcript exclusion**

The second stage involved filtering based on transcript availability, as not all recorded calls were accompanied by transcripts. We only retained calls with available transcripts for analysis.

3. **Third stage: Short transcript length exclusion**
   In the third stage, we further refined the dataset by filtering calls based on transcript length. After applying the log transformation, we excluded transcripts with log values in the lowest 10th percentile of the distribution. Consequently, we exclude calls with excessively short transcripts, which typically lack meaningful content for our analytical purposes, and calls that are unsuccessfully logged, transcribed, and answered. For instance, a call transcript containing minimal content, such as only containing transcripts of “Okay, yeah, yeah.." was considered unhelpful for our analysis and removed. Such calls comprised a mere 0.5% of the text dataset consisting of calls containing words ranging from 1 to 31 words.

4. **Fourth stage: New patient exclusion**
   In stage four, we excluded newly called patients (last call outcome: "Follow-Up Needed") contacted within the final two weeks of the study period. This ensured sufficient time for enrollment consideration after the initial call. A two-week window was chosen to exceed the company’s standard minimum call interval (one week), allowing for increased enrollment opportunities. We found 7 patients who met this criterion and excluded them from the analysis.

Ultimately, this four-stage filtering process resulted in a dataset of 6,741 call IDs, representing 72% of the initial call volume of 9,382 call IDs.

Call transcript

The agent has a predefined script comprised of three sections, including an introduction, a program description, and a program enrollment offer. This document is available in the Appendix 8.6. In the introduction, the patient is informed that they received the call based on their provider’s recommendation. Typically, the agent will add some small talk in this section. Then, the conversation follows with an explanation about the CC program that is delivered through phone calls and its overall benefit. In the event that the patient inquires about insurance coverage, relevant information regarding the patient’s coverage will be conveyed. Lastly, the agent will offer enrollment to the program. Should the patient express consent, prompt scheduling for the coaching process will be organized, while in the event of dissent, the agent will kindly request the patient’s motivation for their decision.

Note that each patient could receive multiple calls before they make an enrollment decision. To incorporate all calls, we sequentially organize the call history of individual patients according to call dates and consolidate all call transcripts into a unified transcript document for each patient. Additionally, we mark the most recent call outcome in the call history as the patient's overall outcome, categorized as either “enroll" or “not enroll," following the categories described in the previous section.

Call metadata

Apart from the transcript, we also include call metadata as features for our enrollment prediction task. Call metadata refers to information about the call. By including the metadata, we could identify whether factors beyond the call's textual content alone hold sufficient predictive power or if the content also plays a significant role in shaping predictions. In this study, we utilize the following five types of metadata available to us:

1. **Total transcript length**. This is the cumulative word count from all transcripts in the patient's call history. Henceforth in the document, when we refer to the patient's transcript, we are referring to the combined transcript of all calls in the patient's history. Total transcript length ranged from 32 to 15,000 words per patient.

2. **Number of calls**. The number of calls denotes the total number of calls received by patients, including voicemails. This count spans from 0 to 14 calls per patient.

3. **Interval**. Interval refers to the gap in days between the most recent call received by the patient and the penultimate call within the enrollment period. The interval values range from 0 to 500 days. A value of 0 means that the patient either received two calls on the same day or only received one call.

4. **Maximum of agent sentiment**. In this study, we use the overall sentiment value, which is automatically provided by Amazon Web Services (AWS). The overall sentiment is derived by averaging sentiment scores across individual responses per speaker within a given conversation. These overall sentiment scores range from -5 to 5. Our analysis focuses on the highest sentiment score expressed by the agent throughout calls with a particular patient. This approach allows us to assess the level of agent sentiment positivity associated with patient enrollment.

5. **Minimum of patient sentiment**. This is the minimum overall sentiment conveyed by the patient across calls, which also ranges from -5 to 5. We only consider the minimum of

the patient's overall sentiment throughout the calls received by the patient to assess the level of negativity associated with patient enrollment.

After conducting an analysis of other metadata, such as the percentage of follow-up and answered calls, we discovered that these other factors did not have as significant a predictive impact as the elements chosen.

To ensure we capture the complete context of the call, we considered transcript length, number of calls, and interval. Additionally, we took into account the highest sentiment of the agent and the lowest sentiment of the patient to gauge the level of positivity conveyed by the agent's words and the degree of negativity conveyed by the patient's words, respectively. These factors will be used as call metadata in the prediction models.

4.2.2 Topic Modeling

Text preprocessing

As mentioned, we focus on word-level analysis to mitigate the challenges of processing a call transcript which is a non-continuous text. We followed a typical process of text preprocessing for each patient's transcript documents as specified in Section 2.1. The process includes removing any URLs and punctuation, converting the text into lowercase, removing stop words, lemmatizing the words, and tokenizing the text into individual words. Additionally, we also removed redaction tags and corrected miss-transcribed words, such as "ice cream" into "eye screen". This approach ensures that we only consider meaningful words for our bag-of-words approach for topic modeling.

Structural Topic modeling (STM)

In this study, we used structural topic modeling (STM) as our topic modeling method [85] as specified in Section 2.3. In this study, the metadata used is the patient enrollment status (i.e., enroll or not enroll). We ran the STM to generate topics from the transcripts that have gone through text preprocessing.

Determining the optimal number of topics to be modeled is a challenging and unresolved issue in topic modeling analysis. Roberts, Stewart, and Airoldi [85] consider the number of topics based on the combination of different measures, such as semantic coherence and word exclusivity within the topic group. However, an optimal number of topics does not necessarily translate to good predictors. Thus, we generated a series of models using 3 to 15 topics for comparison. Each patient transcript has some proportion of topics for each topic model. For instance, for a 3-topic model, a patient's transcript will have some proportion of topic 1, topic 2, and topic 3. Whereas in a 5-topic model, a patient transcript will have some proportion of topics 1 through topic 5, which sum up to one. We used these proportions as a vector representation for each patient transcript. These vectors are referred to as topic proportions and serve as features in the prediction models to predict enrollment.

We employed the FREX (Frequency and Exclusivity) metric to measure the exclusivity of words within topics. FREX takes into account both the frequency of a term within a specific topic and its expected frequency across all other topics, thereby determining its level of exclusivity. It is calculated as the weighted harmonic mean of a word's rank concerning exclusivity and frequency. The next section will explain the topic interpretation process based on their respective topic proportions and these FREX words.

Topic interpretation

Generally, topic models, including STM, do not directly draw conclusions or present specific meanings of topics. Rather, the resulting topics represent groups of words frequently appearing together in documents. These word clusters enable humans to interpret them as topics based on their shared thematic associations. The ultimate goal of this interpretative process is to assign a representative label to each word cluster generated by topic modeling. This is accomplished by examining the FREX words within each topic and the text examples that best illustrate the respective theme (i.e., documents with a high topic proportion of a particular topic).

4.2.3 Prediction models

We developed three models to predict enrollment outcomes in the CC program (i.e., enrollment or not in the program). It is important to note that these models were not designed to predict the outcome of the next call but rather eventual enrollment based on the full call history so far. The first model relied solely on call metadata (metadata model), while the second model used topics derived from STM (topic-based model). The third model combined call metadata with extracted topics (topic + metadata model). We standardized the metadata and topic proportion values so that the mean is zero and the standard deviation is 1, following the assumption made in Tibshirani [103]. The standardization of variables was implemented to render them comparable by mitigating differences in scale and units. We constructed each model using penalized logistic regression with the least absolute shrinkage and selection operator (Lasso), allowing variable selection during model fitting. We evaluated individual predictors and interactions to improve prediction accuracy. Finally, we internally validated the models with 10-fold cross-validation and evaluated their performance based on the area under the receiver operating characteristic curve (AUC) on the test data.

All analyses were conducted using the R statistical software environment, with text processing performed using the “quanteda" package by Benoit et al. [7] and STM implementation carried out using the “stm" package by Roberts et al. [84]. Lasso and cross-validation were performed using the “caret" package by Friedman et al. [34] with “glmnet" method with $\alpha = 1$ (which indicates we use lasso regression).

4.2.4 Evaluation

In this study, we aim to evaluate and compare the performance of several prediction models on a single dataset. Our primary goal is to determine whether there are statistically significant differences in their performance metrics. We employ the non-parametric due to potential violations of normality and independence assumptions to compare the models, as noted by Pizarro, Guerrero, and Galindo [80]. Unlike the Critical Difference Diagram (CDD) by Demšar [24], which is designed for comparisons across multiple datasets, the Friedman test is used focuses because it on identifying overall differences in performance rankings within our single dataset. If the Friedman test reveals a statistically significant difference in performance across all models (indicating at least one pair of models differs), we will proceed with a post-hoc test for pairwise comparisons. The Wilcoxon signed-rank test with Bonferroni correction is a suitable choice for this purpose. This test allows us to identify which specific pairs of models exhibit statistically significant differences in performance while accounting for the multiple comparisons problem arising from testing all possible pairs.

4.3 Results

We obtained data on 6,741 calls from 5,691 patients, following the specified inclusion criteria. The call history was arranged chronologically, and the outcome of the last call for each patient was recorded as either “enroll" or “not enroll". This resulted in a balanced dataset, with 49.81% of patients classified as “enroll" and 50.18% as “not enroll".

Table 4.2 provides a summary of the call metadata values. The median aggregate transcript length for each patient is 815.5 words, with an interquartile range (IQR) of [490 - 1280] (standard deviation (SD) = 861.58 words). Each patient received a median of 2.0 [IQR: 1 - 3] calls, with a standard deviation of 1.77. The median interval between the last call and the call before it is 0 days IQR [0 - 9.97] (SD: 62.64 days). Regarding sentiment, the median score of the minimum patient sentiment across the call history for each patient is 0.5 IQR [0 -1.1] (SD: 0.97) and the median of the maximum agent sentiment is 1.8 IQR [1.1 - 2.5] (SD: 1.00). Finally, all values were normalized using standard normalization to enable comparison between factors.

Table 4.2: The summary of the call metadata variables

Call metadata (unit)	Median [IQR]			SD
	Enroll	Not Enroll	All	
Total transcript length (words)	1,054 [1,285 - 747]	569 [936 - 328]	818 [490 - 1280]	861.58
Number of calls (calls)	2 [1 - 3]	2 [1 - 4]	2.0 [1 - 3]	1.77
Interval (days)	0 [0 - 7]	2 [0 - 11.9]	0 [0 - 9.97]	62.64
Minimum of customer sentiment	0.5 [0 - 1.1]	0.5 [0 - 1.2]	0.5 [0 - 1.1]	0.97
Maximum of agent sentiment	1.8 [1.1 - 2.4]	1.9 [1.1 -2.6]	1.8 [1.1 - 2.5]	1.00

Extracted topics

Through the analysis of the call history of 5,698 patients, we have generated topic models consisting of 3 to 15 topics, assigning a unique mix of topic proportions to each document. For example, a 4-topic model would assign four topic proportions to each document, while a 5-topic model would assign five. These proportions were then utilized as features for classification and standardized to ensure comparability, so that the mean is 0 and the standard deviation is 1 Tibshirani [103].

4.3.1 Predictive performance

Figure 4.2 displays the AUC performance of three model types in predicting enrollment using logistic regression with Lasso on the test set. The metadata model has only one model; the topic-based models consist of 12 models (ranging from 3 to 15 topics), and the metadata+topic models also have 12 models. The number in the circle indicates the number of topics used in the topic model.

The metadata model, which solely uses metadata as a predictor, shows an AUC score of 0.79. The circle is labeled "0" because the metadata model does not use any topics. The topic-based models, which include only topic proportions as predictors, exhibit AUC scores ranging from 0.56 to 0.96. Only the models with 3- and 4-topics obtained AUC scores of less than 0.60, whereas the other models (5-15 topics) had higher AUC scores ranging from 0.93 - 0.96.

In comparison, the AUC score in the topic+metadata models was generally higher than in the metadata and topic-based models. The AUC scores for the topic+metadata models using 3- and 4-topics increased to 0.79 and 0.80, respectively, while the AUC scores of the topic+metadata models with 5-15 topics ranged from 0.96 to 0.97. Figure 4.3 shows the detailed AUC performance for each topic+metadata model. Because the topic+metadata models have the same pattern across different number of topics and generally have higher AUC performance than the topic-based models, we will focus on the topic+metadata models in the next section.

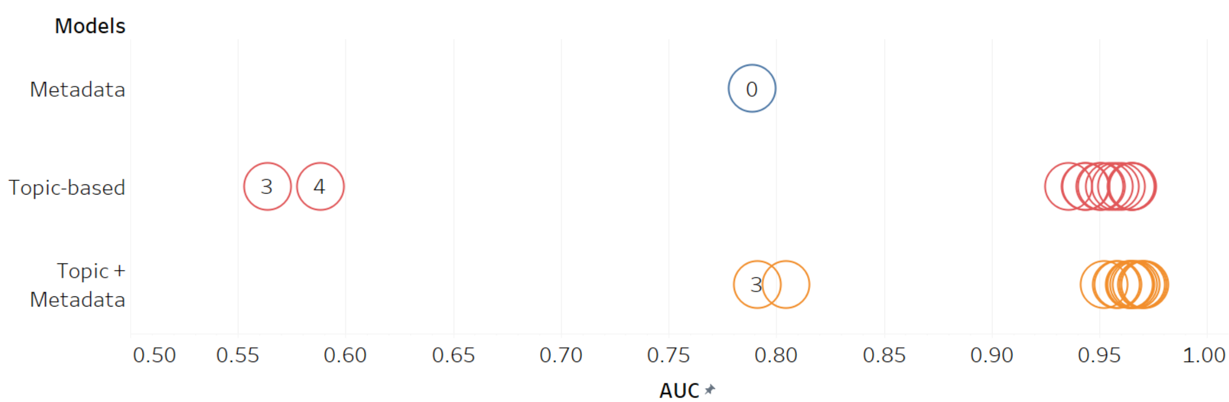


Figure 4.2: AUC performance. The AUC performance of metadata, topic-based, and topic + metadata models for 3 to 15 topic models. The x-axis is the AUC range from 0.5 to 1. Each circle represents each model's performance and the number in the circle indicates the number of topics of each topic model. The number of topics is omitted from the topic model with 5 or more because the circle is overlapped.

4.3.2 Evaluation

We employed a non-parametric Friedman's test to assess potential differences in AUC between the predictive models on the test dataset. We treated topic-based models with different topic numbers as a single model category (i.e., topic-based model) and did the same for topic-

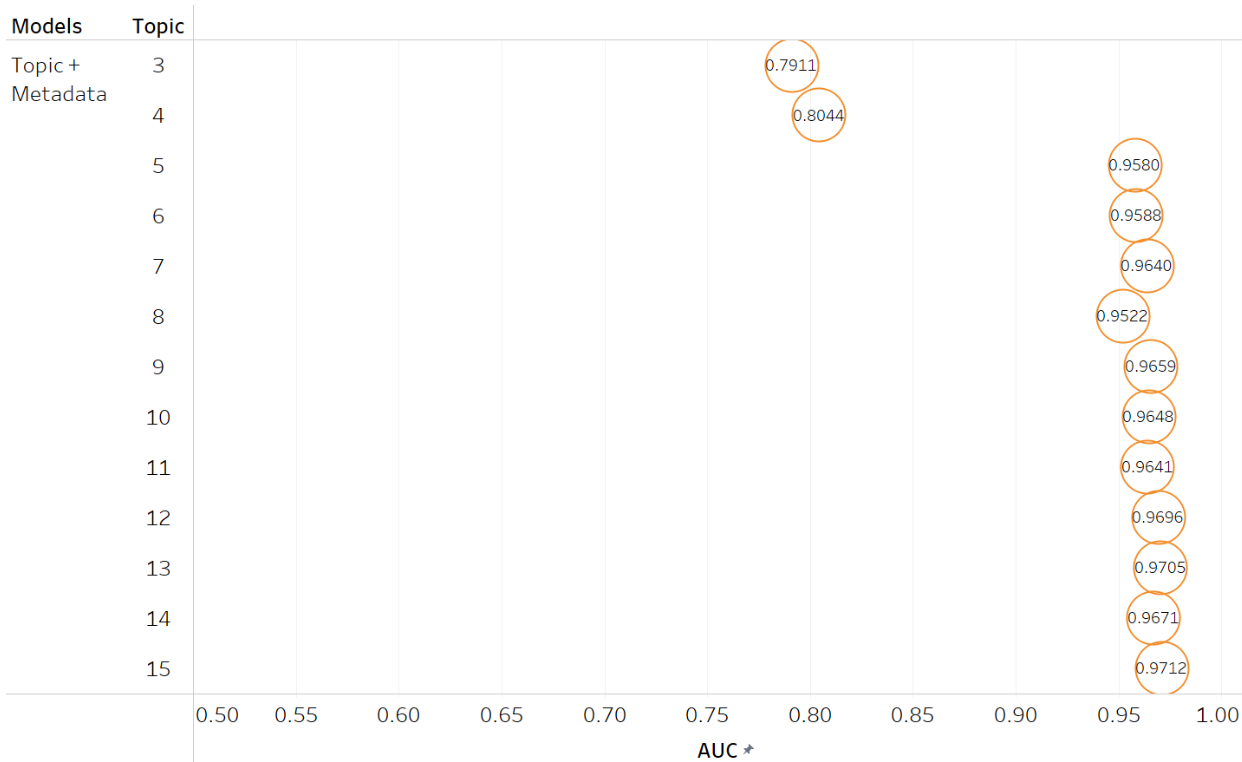


Figure 4.3: Detailed AUC performance of topic + metadata models. The AUC performance of topic + metadata model across 3 to 15 topics. The x-axis is the AUC range from 0.5 to 1. Each circle represents each model's AUC performance.

metadata models. Thus, we treated different numbers of topics as different conditions. This resulted in a total of three model categories for comparison: metadata-only, topic-based (different number of topics considered as different conditions), and topic-metadata models (different number of topics considered as different conditions). The Friedman test revealed a statistically significant difference in ranking ($\chi_2(3) = 22.61$, $p < 0.05$) after Bonferroni correction for multiple comparisons (p_adjusted $= 0.0002$ for all pairwise comparisons). This suggests that all 3 models (i.e., metadata-only, topic-based, and topic+metadata models) likely exhibit variations in their effectiveness on this task.

4.3.3 Call metadata and topics associated with enrollment

In this study, we utilized the odds ratio (OR) to examine the relationship between predictors and outcome status. To account for varying variable scales, we normalized the value and interpreted the OR based on its magnitude rather than its actual value. A higher odds ratio indicates a stronger association between the predictor and outcome.

Our analysis of the metadata model revealed that metadata positively associated with patient enrollment in the program had an OR range of 1.00009 to 1.00130, while negatively associated metadata had an OR range of 0.88237 to 0.99994. We have presented the top five metadata positively and negatively associated with enrollment in Table 4.3 to highlight significant metadata and their interactions. Additional odds ratios for the metadata model can be found in Appendix 8. As per the table, the highest odds ratio belongs to total transcript length (OR 1.00130), while the lowest belongs to the number of calls (OR 0.88237).

As previously mentioned, combining metadata with topic information tends to yield better results than relying solely on topics. As such, we will be focusing on the topic + metadata model. Specifically, we are exploring the 5-topic + metadata model that demonstrated an AUC performance of 0.958, as can be seen in Figure 4.3. We intentionally selected this model due to its significant improvement in performance from 0.804 (as seen in the 4-topic + metadata model) to 0.96. The comprehensive OR scores for the models with different number of topics can be found in the Appendix 8.

Within the 5-topic + metadata model, we discovered that the OR ranges from 2.53 to 153.084 for variables positively associated with enrollment and from 0.28 to 0.59 for features negatively associated with enrollment. Table 4.4 displays the top and bottom five variables and their interactions with the highest and lowest OR values. It is interesting to note that Topic 5.5 (which is the fifth topic in the 5-topic + metadata model) has the highest odds ratio, while the interaction between Topic 5.4 & Topic 5.1 has the lowest. The complete odds ratio value for the 5-topic + metadata model is available in the Appendix 8. Notably, the length of the transcript remains consistent among the top five, and the number of calls is consistent among the bottom five - just as we observed with the metadata-only model earlier.

4.3.4 Topic interpretation for 5-topic + metadata model

As an example of how to interpret the topics, we conducted an analysis of the 5-topic + metadata model, utilizing transcript examples with high topic proportions and FREX words. Note that the interpretation of the five topics in both the 5 topic-based model and the 5-topic + metadata model is the same because the same topics are used. Table 4.5 presents the corresponding topic number, interpretation, top 15 FREX words, interpretation description, and transcript content examples.

Table 4.3: The top five variables exhibiting the highest coefficients and odds ratios, and the bottom five variables displaying the lowest coefficients and odds ratios within the metadata model.

Predictors	Coefficient	Odds Ratio (OR)
Top 5		
Total transcript length	1.274	3.576
Total transcript length & Minimum of patient sentiment	0.365	1.440
Minimum of patient sentiment	0.249	1.283
Interval	0.086	1.090
Interval & Minimum of patient sentiment	0.033	1.033
Bottom 5		
Total transcript length & Interval	- 0.094	0.910
Total transcript length & Maximum of agent sentiment	- 0.102	0.902
Total transcript length & Number of calls	- 0.126	0.882
Maximum of agent sentiment	- 0.169	0.844
Number of calls	- 0.459	0.632

Table 4.4: The top five variables exhibiting the highest coefficients and odds ratios, and the bottom five variables displaying the lowest coefficients and odds ratios within the 5-topic + metadata model.

Predictors	Coefficient	Odds Ratio (OR)
Top 5		
Topic 5.5[a]	5.031	153.084
Topic 5.1 & Topic 5.5	1.777	5.915
Total transcript length	1.397	4.044
Topic 5.4 & Topic 5.5	1.290	3.634
Topic 5.3 & Topic 5.5	0.927	2.527
Bottom 5		
Number of calls	-0.521	0.594
Topic 5.2 & Topic 5.3	-0.635	0.530
Topic 5.4 & Topic 5.3	-0.822	0.440
Topic 5.1 & Topic 5.3	-0.898	0.407
Topic 5.4 & Topic 5.1	-1.273	0.280

[a] Topic 5.5 indicates that this is topic 5 in the 5-topic + metadata model.

Table 4.5: Topic interpretation for the 5-topic model.

Topic number	Topic name	FREX words (top 15-words)	Description	Examples
5.1[a]	Greetings & insurance inquiry	wonder, awesome, helpful, manager, certainly, share, short, purpose, surprise, copay, behalf, healthy, particular, useful, familiar	The initial patient greeting may vary in its content and tone during the conversation.	Agent (A): "Hello, this is calling on **behalf** of doctor [proper noun] ".. "how are you doing today?." Patient (P): "Pretty good..," A:"**Awesome**. Good to hear, doctor had wanted me to reach out just to **share** some information with you about a resource that she's recommending"...
5.2	Patient's overall condition	eat, water, light, kid, dog, body, food, breakfast, cute, heart, sit, die, sick, leg, wow	Inquiry about the patient's current health status and dietary habits. For instance, patients explain how their child primarily assists them in their daily life.	A: "..what's your diet like take me through a day in your diet? ..." P : "... And so we try to **eat** a really good balanced dinner and, and uh so I mean a good no fried **food**s, no fatty, I can't do a lot of that..."
5.3	Patient's medication & diabetic condition	alrighty, grid, vitamin, sign, symptom, check, everything's, [company's name], everything, opt, difficulty, extension, trouble, stable, last	Inquiring about the patient's medication regimen pertaining to their diabetic condition.	P: "For checking my, what, what is that?." A: "Your **grid**? So it's a little **grid** with a dot in the middle that doctor wants you to look at each day to make sure your lines are still straight and there's no curve lines..."

Table 4.5 (continued).

Topic number	Topic name	FREX words (top 15-words)	Description	Examples
5.4	CC program explanation	provide, program, partner, basically, extra, additional, support, flexible, greet, care coordinator, offer, participate, frequent, condition, benefit	Part of the agent's script explains and offers the program to the patient.	A: ... " Uh so this care coordinator is a certified professional who **partner**s with doctor [PII] to **provide** you **extra support** over the phone.."
5.5	Setting up a call appointment	week, text, afternoon, morning, reminder, day, remind, o'clock, around, early, prefer, calendar, technician, cell, message	The words pertain to the timing of the initial coaching call, scheduled once the patient has consented to participate in the program.	A: "I have **morning, afternoon** and I have early evening as well on those two days" P: "I can go into the next **week** if those two days don't work..."

[a] Topic 5.1 indicates that this is topic 1 in the 5-topics model.

4.4 Discussion

This study examined transcripts of calls made during the enrollment process to identify conversation topics and call metadata that may encourage patients to enroll in a telephone-based counseling program, CC program. We found that the identified topics allow further understanding of the factors that lead to enrolling in the program. In addition, we observed that predictive performance could be increased by adding information extracted from the call transcript (unstructured data) to the call metadata (structured data).

Previous research suggests that combining unstructured and structured data can improve prediction performance. Ghassemi et al. [38], Feller et al. [33], and Dormosh et al. [28] found that incorporating unstructured data (i.e., text data) with structured data from electronic medical records (EMRs) such as age, sex, and medication can enhance predictive performance. Our results further support this finding. We found that incorporating unstructured data alongside structured data (i.e., topic + metadata model) can increase a model's predictive performance by providing additional information that structured data may have overlooked. Unlike previous studies, our study utilized feature engineering to obtain our metadata as our structured data.

We further test the three prediction models introduced in this study (i.e., metadata, topic-based, and topic+metadata models) with the Friedman test. The non-parametric Friedman test revealed a statistically significant difference in the AUC performance rankings of the three models ($\chi^2(3) = 22.61$, $p < 0.05$). This finding was further confirmed with Bonferroni correction for multiple comparisons (p_adjusted = 0.0002 for all pairwise comparisons). These results suggest that the models likely exhibit variations in their effectiveness on this task

In this section, we will examine the insights that the metadata and the topic + metadata model provide, specifically for the 5-topic model. As mentioned earlier, we will not cover the topic-only model since the topic + metadata model performs better. We will first explore the call metadata factors that contribute to predicting enrollment. Afterward, we will discuss how topics can be used to support our concluding remarks.

4.4.1 Call metadata

This section focuses on call metadata associated with patient enrollment in a CC program, based on the odds ratio in Table 4.3. The metadata we analyzed includes transcript length, number of calls received, the time between calls (interval), maximum agent sentiment, and minimum patient sentiment. Based on our results, we found that patients are more likely to enroll in the program when they engage in longer conversations with agents, which results in a longer transcript length. However, it's important to note that conversations with enrolled patients tend to be longer due to appointment setup discussions that aren't relevant to patients who decline the program offer. Additionally, our findings suggest that the odds of enrollment decrease when transcript length interacts with other metadata, except when it interacts with the patient's sentiment. In other words, the odds for a patient to enroll increase as the patient's sentiment increases. This interaction indicates that it's beneficial for agents to continue interacting with patients for a longer time as long as the patient's sentiment is perceived to be favorable.

Furthermore, our result suggests that calling patients too often (i.e., increasing the "number of calls") has a lower odds of patient enrolling in the program. Additionally, we found that longer time intervals between calls are associated with higher odds of enrollment. It is important to note that the number of calls and the interval between calls influence each other because taking longer breaks between calls translates to fewer calls made in the same timeframe.

Moreover, we found that the sentiments of patients and agents have a the opposite odds. We observed that a higher sentiment expressed by patients (i.e., positive sentiment) has a higher odds for enrollment. Conversely, the sentiment expressed by the agent was found to have a lower odds for enrollment. Despite the odds ratio of patient and agent sentiment featuring in the top five and bottom five, respectively, the odds was comparatively lower than the factors of "total transcript length" and "number of calls".

Interestingly, one key observation that persists from the metadata-based model in topic + metadata models is the consistency of the "total transcript length" and "number of calls" factors. The "total transcript length" consistently appears in the top five which suggest having high odds for enrollment, while the "number of calls" appears in the bottom five across different predictive models with different numbers of topics. This consistent pattern highlights the significance of these variables.

4.4.2 Topics

Our findings suggest that the identified topics from the NLP method, STM, can provide further insights about patients' enrollment in addition to the structured data (i.e., call metadata). Topics notably enhance predictive performance in topic-based and topic+metadata models, compared to models relying solely on call metadata. To illustrate, we will discuss how the topics identified from the 5-topic model, along with call metadata, were associated with patient enrollment.

By analyzing conversation topics from the STM, we found that topics can provide extended insight into the factors that encourage patients to enroll. Specifically, we found that agents who offer to schedule the first coaching call appointment (Topic 5.5) have higher odds of patient enrollment than those who focus on other topics. This is unsurprising, as patients who enroll will require assistance from the agent to schedule their appointment. However, we were intrigued to discover that the interaction between the topic associated with initial greeting (Topic 5.1) and the topic centered around explaining the CC program (Topic 5.4) had low odds for patient enrollment unless agents also covered the topic of scheduling an appointment (Topic 5.5) and allowed for a longer interval between calls. This outcome was unexpected, given that agents are trained to greet the patient and provide comprehensive explanations to facilitate informed decision-making for patients.

To further understand the unexpected outcome, we conducted a detailed analysis of the topic associated with initial greetings (Topic 5.1) and the topic associated with program explanation (Topic 5.4). As previously mentioned, we found that agents who not only greet (Topic 5.1) and provide patients with information about the program (Topic 5.4) but also offer to schedule an appointment for the first telephone coaching (Topic 5.5) have higher odds of enrolling patients. These interactions indicate that offering to schedule an appointment after greeting the patients and giving a brief explanation about the program has high odds for patient enrollment, even if the patient has not yet decided to enroll in the program. An example of an agent extending an offer to schedule shortly after explaining the program explanation is shown in Figure 4.4. In this conversation, we observed that the agent intentionally offered to schedule an appointment after they explained the program. While it is possible that patients may enroll due to their interest after receiving a detailed explanation, offering to schedule an appointment appears to also have higher odds for patients to enroll. Program explanation by itself may not be enough to persuade patients to enroll. Additionally, we found that agents who explain the program and have longer intervals between calls have higher odds of patient enrollment than agents who call patients in shorter intervals. This implies that the agent can try to space out the calls, allowing the patient some time to contemplate the program and reconsider their decision. Patients may not appreciate receiving a call shortly after one another.

In summary, incorporating conversation topics can significantly improve the predictive accuracy of models when compared to those that solely rely on call metadata. Moreover, it was observed that extended engagement with patients has a high odds in enrollment, while calling a patient too often has a low odds . Furthermore, discussing appointment scheduling during the call has a high odds for enrollment, whereas greeting and explaining the program during the call has a low odds. However, interestingly, if the agent offers to schedule the first appointment with a care coordinator during the conversation, it can increase the odds of patient enrollment.

4.5 Limitation and Future Research

This study has some noteworthy limitations. First, we assumed the call transcripts contained reasons for enrollment in the program. However, external factors beyond the scope of the

Agent : "... You can stop it at any time. Um it is covered under your insurance, subject to your annual deductible coinsurance. Um the billing is all done through [proper noun] I care (eye care) so they would bill your insurance is um but doctor [PII] believe so much in doing this service that he has waived any copays or coinsurance if any were to come out at this time.."

Patient : "Mhm. Yeah..,"

Agent :"So um what I would do today is [last name] you with a care coordinator who will reach out to you for a phone call, initial call about ## or ## minutes of your time and it's based on when you're available um and then they go over the treatment plan and create a care plan for you and then you can decide how often they reach out to you..",Okay. Okay..,Okay. Alright so are you um available most uh most days of the week?."

Figure 4.4: A transcript excerpt of an agent elaborating on a program and subsequently extending an offer to the patient.

calls, such as personal circumstances, may exert influence on patients' decisions regarding enrollment. The external factors could introduce complexity in determining why patients decline the offer during the call, and even when agents inquire for explanations, patients are not obligated to provide detailed responses. Moreover, patients may harbor concerns that the calls are phishing attempts, potentially explaining their hesitance to engage further. It's noteworthy that the company has taken steps to address this by implementing caller ID, aiming to instill greater confidence in patients regarding the authenticity of the calls.

Second, the STM framework has limitations in capturing semantic nuances, synonyms, and antonyms due to its reliance on word occurrence. This poses a challenge for the framework in identifying phenomena such as the dual meanings of certain words like "may." However, this issue is not unique to STM but is prevalent in another topic model without word embedding [19].

Future research could address these topic model limitations by exploring alternatives such as word embeddings, neural topic models, or models considering document structure, like Paragraph Vectors. These methods offer the potential for capturing richer semantic and contextual information in text data. In the context of call transcripts, developing auto-removal of the agent's program explanation or scheduling the first appointment in call transcripts can improve topic extraction.

4.6 Conclusion

The primary goal of this research is to investigate the factors that lead patients to enroll in a telephone-based counseling initiative (i.e., the care coordination program). We analyzed the transcripts of the calls made during the enrollment process and transformed them into vector representations using an NLP method known as Structural Topic Modeling (STM). This method extracts a mixture of topics from each patient's transcripts as a vector representation. We used these topics, along with call metadata, as transcripts' features, to predict patient enrollment. We compared the predictive performance of three models, call metadata, topic-based, and combination of topic and metadata (i.e., topic+metadata) model.

Out of the three models, the topic + metadata model outperforms the other two models in distinguishing between patient enrollment and non-enrollment. We further use the Friedman test to identify the overall differences in the AUC performance ranking. The test revealed significant differences in AUC performance rankings among the three models (i.e., metadata, topic-based, and topic + metadata) with $\chi^2(3) = 22.61$ and $p < 0.05$, confirmed by Bonferroni correction for multiple comparisons (p_adjusted = 0.0002). This indicates variations in model effectiveness for the prediction task. By incorporating topic features from unstructured data to the call metadata as structured data, the predictive capability of the model can be enhanced. Furthermore, adding topics to the prediction model can increase the model's comprehension of the relationship between unstructured data and factors associated with outcome prediction (i.e., patient enrollment).

We found both expected and unexpected results from our topic + metadata model example (i.e., 5-topic model). It is expected that the topic related to scheduling telephone counseling appointments (Topic 5.5) had a high odds for patient enrollment. However, we did not expect that conversations regarding greetings (Topic 5.1) and program explanations (Topic 5.4) would have low odds for enrollment. Interestingly, agents who offer to schedule an appointment for telephone coaching have higher odds of enrolling patients than just greeting or explaining the program. Greetings and explaining the program in detail may not be enough to persuade patients to enroll. Therefore, offering to schedule an appointment appears to have high odds. Furthermore, agents who engage in prolonged conversations tend to have higher odds of enrolling patients across different models. Conversely, calling patients has low odds, which is consistently shown across different models.

The CC program aims to offer patients extensive assistance in addressing their queries related to diabetic retinopathy. The findings in this study provide insights to agents in the enrollment process to evaluate their strategy in calling patients, which in turn can increase program participation among patients. The agent in the enrollment process is the first contact point between patients and the CC program, which is a critical phase in determining whether the patient enrolls and eventually accesses the benefits of the program. The success of an agent to enroll a patient in the program is determined not only by external factors but also by the conversation that takes place during the call. This study attempts to capture this by analyzing the conversation during the call to draw insights. Therefore, this research has far-reaching implications for improving the accessibility and effectiveness of the CC program in promoting better outcomes for patients with DR.

CHAPTER

5

EXTRACTIVE TEXT SUMMARIZATION IN QUALITATIVE RESEARCH: A TOOL FOR CONTENT ANALYSTS

5.1 Introduction

The exponential growth of text data in healthcare, including biomedical literature and electronic health records (EHRs), presents significant challenges for information access, management, and knowledge extraction. Extracting relevant information from this vast and ever-growing data pool is a time-consuming task that can hinder optimal patient care [110]. Biomedical text summarization (BTS) has emerged as a promising solution to address this challenge, offering clinicians and researchers tools to efficiently synthesize information and reduce the time needed to stay current with the latest advancements [102].

BTS aims to generate concise and informative summaries that capture the key points from various biomedical sources, including journal articles, clinical trials, and patient records [110, 102, 71]. These summaries can serve diverse purposes, such as providing researchers with updates in their field [21, 78], supporting clinicians in staying abreast of evidence-based practices, and facilitating time-saving summarization of clinical data like radiology reports [52, 95, 65] and patient notes [79, 32]. By rapidly comprehending patient-specific information and relevant evidence, BTS empowers clinicians to make informed clinical decisions and deliver optimal, continuous care [110, 102].

Within BTS, Query-focused Multi-Document Summarization (QMDS) holds particular promise for healthcare. QMDS allows users to pose specific questions and receive summaries that directly address their information needs [86, 5]. This functionality is valuable in navigating the vast amount of information healthcare professionals encounter [72]. To generate this QMDS summary, BTS can be categorized into three main approaches: extractive, abstractive, and hybrid text summarization [56]. Extractive methods select key sentences from the source text. Abstractive methods generate entirely new summaries capturing the main ideas. Hybrid methods combine both approaches

Recent advancements in large language models (LLMs) like GPT-4 [76] have garnered significant attention for their abstractive summarization capabilities in the biomedical domain [21]. Despite their potential, LLMs in healthcare have challenges, such as intricate medical terminology, variable clinical documentation, and the critical need for exceptional reliability and accuracy [113]. To address these challenges, researchers have explored training LLMs from scratch on biomedical data or fine-tuning existing models with domain-specific datasets [118, 12, 93, 64]. However, some limitations and risks are still present. Tian et al. [102] identified some limitations and risks that require further investigation, which include hallucination, fairness and bias, the risk to patient privacy, and lack of comprehensive evaluation.

Given the identified limitations and risks, healthcare professionals considering LLMs implementation should exercise significant caution. While research exploring ChatGPT's potential applications in biomedicine and clinical settings exists [89, 60], real-world deployments of ChatGPT or any other LLMs in healthcare remain unreported [102]. This cautious approach is necessary due to the high stakes in healthcare, where privacy, reliability, and accuracy are crucial [113]. Thus, it is preferable for LLMs in healthcare to be customizable by leveraging their own confidential data while adhering to rigorous privacy-preserving techniques when implementing the LLMs for text summarization. By leveraging pre-trained models and tailoring them to specific needs or patient characteristics, healthcare institutions can achieve significant advancements in areas like clinical decision support, personalized medicine, and efficient information retrieval from biomedical information, such as electronic health records (EHRs).

To achieve a rigorous privacy-preserving and customizable model, the decision between implementing closed-source and open-source LLMs presents an additional layer of complexity in healthcare. Open-source models like BLOOM [8] offer transparency in their source code, architecture, training data, and mechanism for training and inference, which allow full user control for customization. However, they often lack dedicated support, potentially hinder-ing successful deployment. In contrast, closed-source models like the commercial version of GPT4 from OpenAI [60] and Gemini from Google [40] provide dedicated support and resources. However, this approach offers less transparency into the source code and training data, raising concerns about potential bias within the model and security vulnerabilities associated with submitting data for analysis. In healthcare, limited transparency in training data and source code of closed-source LLMs hinders healthcare professionals' capabilities to gain full insight from their own data. This restricts customization, private data training, and effective bias assessment and performance evaluation. On the other hand, open-source LLMs offer potential advantages in privacy and customization through private data training, but LLMs are susceptible to "hallucination," which can be detrimental in healthcare settings.

Other abstractive methods besides LLMs present another challenge due to their inherent complexity. Most of the abstractive methods are supervised learning and relevance-based approaches, which are complex to build [56]. Although supervised learning methods have achieved impressive results, their drawback lies in the substantial amount of training data required and its corresponding human-generated summaries [6].

Extractive summarization (ES), a BTS method, offers an alternative to mitigate risks in abstractive summarization (AS) methods, particularly the risk of generating factually incorrect information ("hallucination") often associated with LLMs. ES directly selects key sentences from source documents and concatenates them, ensuring factual accuracy and preserving authorial intent. However, this method suffers from redundancy, longer extracts, and potential incoherence [56]. Furthermore, as information from multiple resources is available and query-based summary is desirable, dealing with QMDS further amplifies these challenges, requiring the model to maintain a coherent, non-redundant, relevant, factually consistent, and grammatically readable summary [66, 5]. Redundancy and relevancy are particularly noteworthy issues to be considered in QMDS, where maintaining a concise and relevant narrative becomes crucial.

In this study, we propose an ES method that mitigates the risk of redundancy and lack of relevancy. Redundancy arises when the model selects sentences with similar semantics for the final generated summary. Lack of relevancy means the summary is irrelevant to the user's query. The proposed ES method is adapted from a method introduced by Belwal, Rai, and Gupta [6] that attempted to address redundancy by grouping similar sentences and tackling lack of relevancy issues by identifying topics across those sentence groups to ensure topic coverage. However, we diverge from the original method in our sentence selection process. While Belwal's method selected the sentence with the highest similarity score within each cluster, our approach will select the highest-scoring sentence from each cluster and topic. This modification ensures that the final summary incorporates the most relevant sentences across clusters and specific topics within those clusters.

In particular, this study introduces and compares two extractive summarization (ES) methods: unsupervised (UESM) and supervised extractive summarization method (SESM). The UESM focuses on producing a generic summary, while the SESM generates a summary focused on a specific topic (i.e., QMDS). While both methods leverage the same core principles of sentence selection and ranking based on the mentioned adjusted Belwal's method, they diverge in their input document pre-processing steps. In the unsupervised extractive summarization method (UESM), the input documents are individual sentences in the papers that have been parsed from multiple documents. In contrast, the supervised extractive summarization method (SESM) input documents are selected sentences that are considered relevant to a particular topic to be summarized. These pre-selected sentences will be referred to as coded sentences. Sentences are coded by adapting the coding process in the content analysis (CA) method, a qualitative method. Thus, the ES method can act as a complementary tool that assists CA users in identifying key sentences that capture the main themes within the pre-selected pool. Further details of the proposed method are available in section 5.2.4.

As a case study, we attempted to summarize multiple biomedical literature based on a user's query (i.e., QMDS). In particular, we focus on summarizing the trade-offs between providing and promoting colorectal cancer (CRC) screening in the available literature related to the CDC's Colorectal Cancer Control Program (CRCCP). To show that the generated summary from the SESM summarizes specific topics in the documents, we compare the generic summary produced by UESM with the SESM.

The remaining sections are organized as follows. In the next section, we will provide the background of CRCCP and describe the proposed methods. Subsequently, the evaluation results for the method will be presented. Finally, a discussion section will be included to provide critical insights and conclusions.

5.2 Methods

In this section, we will first introduce the CRCCP and the key question guiding our goal of summarizing the selected journal papers on this topic. We will then describe the data used, the query-based multi-document summary (QMDS), and our proposed method.

In particular, this study proposes and compares two ES methods: unsupervised (UESM) and supervised extractive summarization method (SESM). UESM produces a generic summary, while SESM generates query-based summaries from multiple documents (i.e., QMDS). We will compare UESM as the baseline method with SESM to show that the summary generated from SESM focuses on a particular topic (i.e., query-based), not a generic one. While both methods leverage the same core principles of sentence selection and ranking based on the mentioned adjusted Belwal's method, they diverge in their input document pre-processing steps. UESM directly analyzes individual sentences parsed from multiple documents. In contrast, SESM uses pre-selected sentences ("coded sentences") relevant to a specific topic for summarization. These coded sentences are generated through the content analysis (CA) process, a qualitative method. By leveraging pre-selected sentences, SESM acts as a complementary tool for CA users, assisting them in identifying key sentences that capture the main themes within a focused topic.

For the proposed method, we will first describe the UESM, followed by the CA and the SESM, which combines CA and ES. UESM, detailed in Section 5.2.4, targets generic summaries from multiple documents. SESM (Section 5.2.5), leveraging a content analysis (CA) process, focuses on query-based summaries.

Figure 5.1 depicts the overall workflow. We build upon a parallel CA study by our team (details omitted) for ES development and summary evaluation. The section where each method is described is shown in parentheses. Note that our study parallels a CA study conducted by our CA team. We leverage their study to develop our ES and summary evaluation.

5.2.1 The Colorectal Cancer Control Program (CRCCP)

Colorectal Cancer (CRC) originates predominantly from abnormal growth in the colon or rectum and stands among the top 5 cancers with elevated rates of both new cases and fatalities between 2016 and 2020 [43]. According to the U.S. Preventive Services Task Force, individuals aged between 45 and 75 years should undergo regular screening tests for CRC as early detection is the key to preventing and treating cancer [17]. The screening test aims to identify the disease before any symptoms occur since survival rates are higher when CRC is detected before it spreads to other parts of the body [1].

The Colorectal Cancer Control Program (CRCCP) is implemented by the Centers for Disease Control and Prevention (CDC) with the aim of increasing CRC screening uptake among individuals aged between 45 and 75 years. The program especially focuses on clinics that serve lower-income people and have low screening uptake (i.e., clinics where fewer than 60% of patients are up-to-date on screening) [15]. However, despite these efforts, the uptake of colorectal cancer screening remains low among groups that face a disproportionately high burden of colorectal cancer incidence and mortality [87]. According to Healthy People 2030, only 58.7% of the target group received a CRC screening, which is below the target of 68.3% [2].

The Colorectal Cancer Control Program (CRCCP) has undergone changes over the years to maintain its effectiveness [15]. In its initial phase, the program's primary focus was promoting

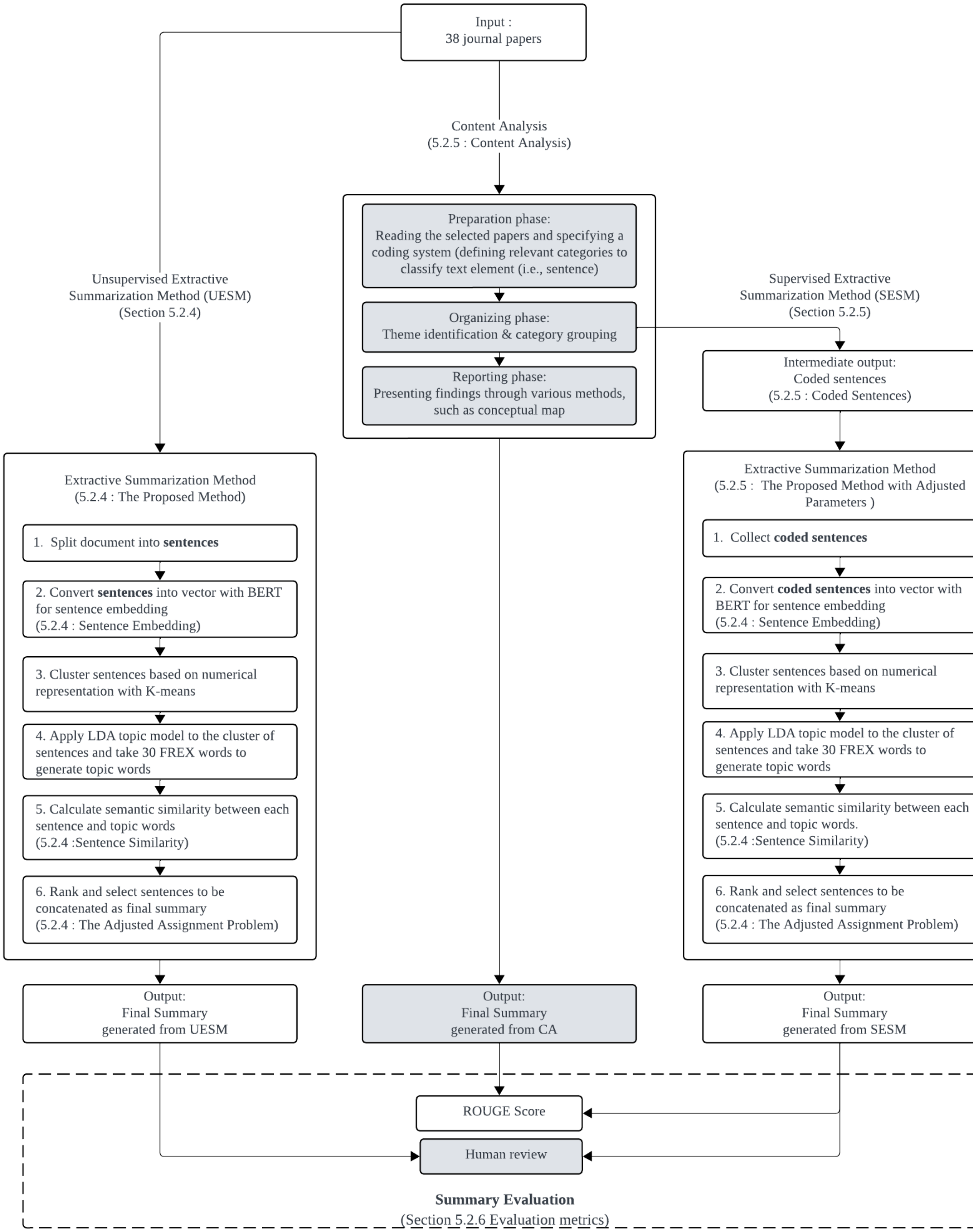


Figure 5.1: Overview of methods presented in Chapter 5.2 with corresponding sections for the different approaches. The grey-shaded processes represent steps performed by the content analysis team.

colorectal cancer screening among individuals over 50 years of age and directly providing them with screening services. However, the program has now made it mandatory for its awardees to partner with clinics that cater to high-need populations and implement interventions based on evidence (i.e., evidence-based interventions (EBIs)) that have been shown to be effective in increasing the rate of colorectal cancer screening. In this program, awardees can be states, universities, tribal organizations, or other organizations. In the current iteration, which began in 2020, the CRCCP continues to support the implementation of EBIs, strongly emphasizing increasing diagnostic colonoscopy completion. The EBIs that have been proven to work include patient reminders, provider reminders, reducing structural barriers, provider assessment and feedback, and patient navigation [16]. Currently, the CRCCP funds 35 awardees, including 20 states, 8 universities, 2 tribal organizations, and 5 other types of organizations.

Several papers have discussed the impacts of changes made to the CRCCP over time. The cost of providing screening and promoting screening has been highlighted by Subramanian et al. [97] and Tangka et al. [99], respectively. Maxwell et al. [67] compared the screening provision and promotion strategies between grantees and non-grantees. Additionally, various studies have explored effective EBIs in promoting screening [22, 91]. However, to the best of our knowledge, no studies have compared the trade-offs between the provision and promotion strategies. To summarize the advantages and disadvantages of the provision and promotion strategies from journal papers that have discussed these strategies separately, our research team conducted a content analysis (CA). This CA effort was conducted in parallel with our study. The process details are available in Section 5.2.5: Content Analysis (CA).

Based on their analysis, the CA team found six major themes related to trade-offs between provision and promotion strategies in CRCCP to increase the screening uptake from the literature. They are “Recommendation for improving screening or testing qualities”,“Efficiency”, “Cost considerations”, “Selecting which EBIs to implement”, “Implementation and effectiveness of EBIs”, and “Strategies for maximizing EBI duration or sustainability”.

In this study, we attempted to identify these types of trade-offs from the literature automatically. To achieve this objective, we propose to use an extractive summarization method, which will be explained further in Section 5.2.4 for UESM and Section 5.2.5 for SESM.

5.2.2 Data

Data Inclusion Strategy

This study aims to summarize the key trade-offs involved in providing versus promoting colorectal cancer colorectal cancer (CRC) screening by CRCCP programs (CRCCP) in increasing the CRC screening uptake. Our research team curated a dataset of 38 relevant journal articles published between 2011 and 2023.

These papers were searched and selected by two researchers who conducted the CA analysis. Both researchers were actively involved in reading the 38 papers manually and highlighting relevant sentences to the topic (i.e., coding the sentences). Following is the search and selection process of the 38 papers:

- Literature Search Strategy:
  - The first reviewer searched PubMed and Google Scholar for:
    - "CRCCP"
    - "Colorectal Cancer Control Program"
    - "colorectal cancer screening provision"
    - "colorectal cancer screening promotion
  - The second reviewer searched for publications by prolific CRCCP authors (e.g., DeGroff, Sharma, Tangka).
- Selection Process:
  - A total of 46 publications mentioning CRCCP were identified
  - Eight articles were excluded: companion summaries or lacking information on CRC screening provision vs. promotion trade-offs.
  - The remaining 38 papers were included for analysis.

Text Data Description

We include all sections in the papers except for the abstract, references, footnotes, and supplementary documents. We include these sections in our data set to be comparable with the CA effort of reading all sections manually to select sentences that are considered to be relevant to the topic at hand. The average paper in the dataset contains 121 sentences (standard deviation: 38.11). The sentences are 27.26 words long on average (standard deviation: 14.05 words). In total, the dataset has 4616 sentences across papers.

5.2.3 Query-based Multi-Document Summarization (QMDS)

The abundance of text data has made it crucial to have the ability to summarize large amounts of information in a concise manner. This phenomenon is also true in biomedical and healthcare information fields [110]. Biomedical text summarization (BTS) is a promising solution to this challenge, as it aims to generate a brief summary representing the key ideas of a document [56, 110, 66]. The main idea behind BTS is to develop computational techniques or algorithms that can condense large volumes of text into shorter, coherent summaries while retaining the essential information and meaning of the original text [66].

BTS can be broadly categorized based on the number of sources, summary triggers, and construction methods [5].

- **Based on the number of sources.** Based on the number of sources, BTS can be categorized into single and multi-document. Single-document summarization (SDS) is relatively simpler to perform, but it may not generate a comprehensive summary as it lacks data [66]. Conversely, multi-document summarization (MDS) is a complicated task involving lengthy input documents written at different times, covering diverse perspectives, and containing redundant information [100]. Therefore, MDS becomes challenging as the method is expected to be able to retain the most critical content while generating a coherent, non-redundant, factually consistent, and grammatically readable summary [66]. Despite the differences, SDS and MDS methods share learning strategies, evaluation metrics, and objective functions.

- **Based on the trigger.** BTS can be categorized as generating generic or query-based summary (i.e., user-focused or topic-focused) [5, 86]. Generic summary means that the summary generated is general, while a query-based summary is based on the user's query or summarizing specific information.

  Existing text summarization mainly focuses on generating generic summaries, and limited research focuses on the query-based one [5, 86]. In addition to the issues mentioned regarding generic summaries (e.g., redundancy), query-based has additional challenges in relevancy and diversity. Relevancy refers to how well information pertains to a specific topic or user query. While selecting the most important information is the main task for producing a generic summary, the model should also ensure its relevance to a particular topic in generating a query-based summary. Diversity refers to considering different aspects of the question in the summary to the greatest extent possible to ensure user satisfaction. This challenge is difficult to overcome since the model should understand the connections between sentences to the query [5].

- **Based on the methods.** There are three methods: extractive, abstractive, and hybrid [56]. Extractive methods identify and select key sentences from the source text, while abstractive methods generate entirely new summaries capturing the main ideas. Hybrid methods combine both approaches.

Each method has its own advantages and disadvantages. Most abstractive techniques (e.g., supervised learning-based and relevance-based methods) are challenging to implement due to their complexity [6, 56]. While supervised learning-based methods have shown excellent performance, their primary limitation is that they require a large amount of training data in the form of input documents and their respective summary examples [6, 66]. The extractive approach is faster and easier to implement than the abstractive approach, and readers can read the summary with the same terminologies that exist in the original text [98]. However, the approach faces challenges, such as redundancy, longer sentences extracted, conflicted expressions, lack of semantics and cohesion, and crucial information being dispersed among sentences [56]. The hybrid approach can improve the overall summarization performance by complementing the weaknesses between extractive and abstractive methods [111]. However, the quality of the abstracted summary generated from selected sentences in the extractive summary may be inferior to the summary results from pure abstraction from the original text.

This study will focus on the query-based multi-document summary with ES method to summarize the journal papers. We chose the extractive approach because it preserves the original sentences, ensuring that critical information is retained. This can be crucial in summarizing scientific papers where specific details and findings are essential or certain words carry specific meanings (such as the word "objective" in optimization methods). Moreover, ES can also mitigate the "hallucination" issue in LLMs that are particularly detrimental in healthcare settings [102]. Additionally, the extractive approach is faster and easier to implement than the abstractive approach, where the techniques are complex and require a large number of training (i.e., input documents) and summary examples [6]. Despite being faster and simpler, ES achieves higher accuracy by directly extracting sentences from the original text [98]. This characteristic aligns with our goal of maintaining the meaning and terminology used in the original journal articles.

In general, the ES approach follows this process:

1. Converting text into suitable representation for text analysis. This text can be represented as word vectors using techniques like N-grams, TF-IDF, etc. Alternatively, it can be split into sentences and converted to vector form.[112]

2. Assigning scores to sentences for ranking purposes. Rank sentences based on the input text representation.

3. The Final summary is extracted from concatenating high-scored sentences. It is created by combining sentences with high scores. These sentences are considered important

compared to others in a document or multi-document. The summary length can be adjusted based on the preferred summary length threshold.

5.2.4 Unsupervised Extractive Summarization Method (UESM)

We want to emphasize that the UESM is the baseline method that will be used as comparison to the SESM. In particular, we want to show that SESM generates a query-based summary compared to UESM that produce a generic summary.

The Proposed Method

Building upon the work of Belwal, Rai, and Gupta [6], this study seeks to address redundancy and improve relevancy in extractive summaries (ES). Redundancy, where summaries contain repetitive information, can be tackled through clustering similar sentences, as proposed in Belwal's method. This approach ensures that only one representative sentence is chosen from each cluster, minimizing redundancy. Relevancy, on the other hand, refers to the selection of important content from the resources [57]. We adapted the Belwal's method of identifying topics across the previous sentence clusters to enhance relevancy by ensuring topic coverage.

Belwal's method involves several steps. First, the text is divided into sentences. These sentences are then converted into numeric representations, such as vectors. Next, based on the numerical representations, similar sentences are grouped together with a clustering method such as K-means. topic modeling is then applied to identify topics across the clusters. To select the relevant sentences for the summary, the method calculated the similarity between each sentence and the words associated with each topic from the topic model. Sentences with high similarity scores are then chosen from each cluster and combined to form the final summary.

We employed Belwal's method to generate the generic summary from the journal papers with a couple of adjustments. First, their method was designed to summarize a single document, whereas we applied it to multiple documents. Second, their approach involved ranking sentences within each cluster by similarity score (highest to lowest) and selecting the top-ranked sentences. We adjusted the sentence selection process to prioritize sentences across clusters and topics that maximize the overall similarity score by solving an adjusted assignment problem. This ensures we select informative sentences that minimize redundancy and increase relevancy while covering all relevant topics. The details of this adjusted sentence selection method are explained in Section 5.2.4.

Figure 5.2 illustrates the adapted unsupervised extractive summarization process, building upon Belwal's method. Following is the description of each step.

1. **Split the document into sentences**.

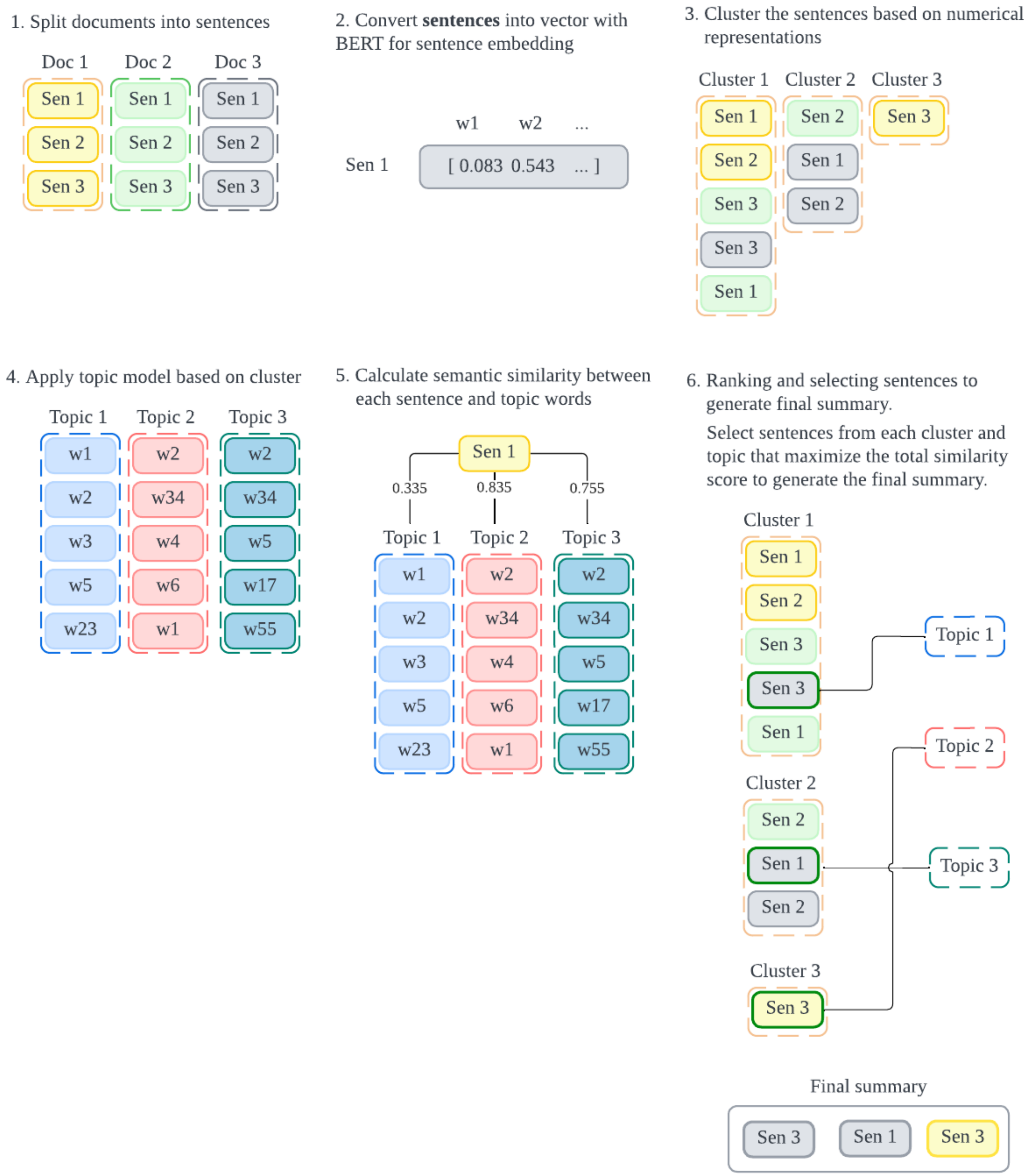


Figure 5.2: Unsupervised Extractive Summarization Method (UESM)

2. **Convert each sentence into a numerical representation**. We convert each sentence into a numerical representation suitable for text analysis tasks. We embedded each sentences with a pre-trained method from Bidirectional Encoder Representations from Transformers (BERT). The details of the BERT used are discussed in the next section (Section 5.2.4: Sentence Embedding)

3. **Divide sentences into different clusters**. We employed the K-means clustering technique and chose the value of K based on the desired number of sentences for the summary as suggested in Belwal, Rai, and Gupta [6]. We generated summaries of 13 and 26 sentences to be reviewed. We based this number on the number of identified subthemes from the CA effort. However, human evaluation favored the conciseness and focus of the 13-sentence summary. Therefore, we will concentrate our further analysis on this length.

4. **Apply the topic modeling on clusters.** This approach treats each document as a cluster of sentences instead of individual sentences. This approach identifies a mixture of topics within each cluster, revealing the thematic distribution at the cluster level instead of focusing on individual sentences. Latent Dirichlet Allocation (LDA) is the chosen topic modeling technique, and the number of topics chosen is 13, similar to the number of clusters previously defined. To represent the thematic content of each topic, we select and concatenate 30 FREX words (referred to hereafter as topic words) for each topic.

5. **Calculate semantic similarity** To identify the most relevant sentences within each cluster, we calculated their semantic similarity with the topic words. Sentence similarity was determined by embedding sentences and topic words using Bidirectional Encoder Representations from Transformers (BERT) and then calculating their cosine similarity. The details of this process are explained in Section 5.2.4: Sentence Similarity. The value ranges from -1 to 1, where -1 indicates perfect negative correlation, and 1 indicates perfect similarity.

6. **Find important sentences for the final summary.** Unlike Belwal's method, which selected sentences with the highest similarity scores in each cluster, we focused on maximizing the total similarity score from sentences chosen in each cluster to at most one topic. This approach ensures a more comprehensive representation of the key themes present. We achieved this by employing a modified version of the assignment problem, which is detailed in Section 5.2.4: The Adjusted Assignment Problem.

Sentence Embedding

Our sentence embedding uses a compressed variant of the Bidirectional Encoder Representations from Transformers (BERT) model from Devlin et al. [26] that is trained using a Pre-trained Distillation training method by [107]. The training method is built on model compression by Buciluă, Caruana, and Niculescu-Mizil [13] and its variant knowledge distillation by Hinton, Vinyals, and Dean [48]. The knowledge distillation technique involves creating a smaller, faster model (student) that mimics the performance of a larger, more powerful model (teacher) in BERT. Using this technique, the Pre-trained Distillation algorithm pre-trains the student model on a large corpus of unlabeled data and then further refines it using distillation from the teacher. As the result of the training method, Turc et al. [107] provides a compact version of 24 BERT models.

This study employs the smallest model called Transformer$_{TINY}$, which has 2 transformer layers (L) and 128 hidden embeddings (H). The model contains 4 million parameters, which is significantly less than the teacher model, BERT$_{LARGE}$, which has 24 L, 1024 H, and 340M parameters. We used Transformer$_{TINY}$ to embed coded sentences, resulting in a numerical representation with 128 dimensions.

Sentence Similarity

We calculate the similarity score between a sentence and topic words by embedding both texts with BERT and calculating the cosine similarity. Specifically, we employed Sentence-BERT by Reimers and Gurevych [82] to generate sentence embeddings. This approach leverages the pre-trained BERT model to create a single vector representation for each sentence, capturing its overall meaning. We specifically used the “all-MiniLM-L6-v2” model, which offers a maximum sequence length of 256 tokens, utilizes 6 transformer layers (L) and possesses 384 hidden embedding dimensions (H). Finally, mean pooling was applied to the sequence of word vectors generated by BERT to obtain the final sentence embedding.

Cosine similarity was then calculated between these sentence and topic word embeddings to assess their semantic alignment. If **u** is a sentence embedding and **v** is a topic words embedding with n-dimensional vectors, then the cosine similarity is calculated as follows.

$$\cos(\varphi) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|} = \frac{\sum_{i=1}^{n} u_i v_j}{\sqrt{\sum_{i=1}^{n} u_i^2}\sqrt{\sum_{j=1}^{n} v_j^2}} \tag{5.1}$$

The Adjusted Assignment Problem

This section attempted to address the challenge of assigning a sentence within a cluster to at most one topic. This ensures the selection of the most relevant sentences from each topic, minimizing redundancy and maximizing topic coverage. We formulate this problem as an optimization model aiming to maximize sentence-topic similarity while imposing the constraint that each cluster can only contribute one sentence to a particular topic. To mathematically model this problem as an optimization problem, we will define the following notation and formulate the objective function and constraints.

Notation:
$I = \{1,..,m\}$ be cluster of the sentences
$J = \{1,..,m\}$ be topics
$K = \{1,..,o\}$ be sentences

Parameters
s_{jk} = be the similarity score of sentence k with words in topic j
K_i = be subset of K such that sentence k in cluster i, $K = \{K_i\}_{i\in I}$

Then the decision variable,

$$x_{ij} = \begin{cases} 1 & \text{if cluster } i \in I \text{ is assigned to topic } j \in J \\ 0 & \text{otherwise} \end{cases}$$

$$y_{jk} = \begin{cases} 1 & \text{if sentence } k \in K_i \text{ is assigned to topic } j \in J \\ 0 & \text{otherwise} \end{cases}$$

Given this definition, the problem formulation is

$$\text{Maximize} \quad \sum_{j\in J}\sum_{k\in K} s_{jk} y_{jk} \tag{5.2}$$

$$\text{Subject to:} \quad \sum_{k\in K_i} y_{jk} \leq x_{ij} \qquad \forall\, i\, j \tag{5.3}$$

$$\sum_{i\in I} x_{ij} \leq 1 \qquad \text{for } j = 1,\ldots,m \tag{5.4}$$

$$\sum_{j\in J} x_{ij} \leq 1 \qquad \text{for } i = 1,\ldots,m \tag{5.5}$$

$$x_{ij}, y_{jk} \in \{0,1\} \tag{5.6}$$

Following are the explanations for each constraint:

- The objective function (5.2) tries to find the assignment of sentences to topics that maximizes the total sum of similarity scores.
- Constraint 5.3: This constraint limits the number of sentences assigned to a topic from a specific cluster. It states that the sum of y_{jk} (assignments of sentence k in cluster K_i to topic j) must be less than or equal to x_{ij} (which indicates whether cluster i is assigned to topic j or not). This ensures that a sentence can only be assigned to topic j if the entire cluster i (where the sentence belongs) is assigned to that topic j.
- Constraint 5.4: This constraint limits the number of topics a cluster can be assigned to. It states that the sum of x_{ij} (assignments of clusters to topic j) must be less than or equal to 1 for each topic j. This ensures a cluster can only be assigned to at most one topic.
- Constraint 3 (Equation 5.5): This constraint limits the number of clusters assigned to a topic. It states that the sum of x_{ij} (assignments of clusters to topic j) must be less than or equal to 1 for each cluster i. This ensures a topic can only be assigned to at most one cluster.

We solve this problem by adjusting the Hungarian algorithm to accommodate the assignment of sentences in each cluster to each topic. The Hungarian algorithm or Kuhn-Munkres algorithm from Munkres [74] is a combinatorial optimization algorithm used to solve the assignment problem, which is assigning a set of agents to a set of tasks so that the total cost is minimized. Let the rows be the list of agents, the columns be the list of tasks, and the matrix be the cost of assigning one agent to any task. The algorithm starts by subtracting minimum values from rows and columns to find the minimum values in rows and columns. Then, it seeks to cover all zeros in the matrix with minimal horizontal and vertical lines, indicating potential low-cost assignments. The algorithm stops when it can cover all zeros in the cost matrix using exactly the same number of horizontal and vertical lines as the number of rows or columns (whichever is smaller). This ensures that every agent is assigned to a task, and every task has an assigned agent, with minimal total cost.

Building upon this algorithm, we adjust a couple of aspects. First, we redefine rows to represent groups of sentences within a cluster rather than individual sentences. Consequently, the rows in the cost matrix will correspond to the sentences in cluster i, the columns will represent the set of topics, and each element in the matrix will reflect the similarity score between sentence k in cluster i and a word topic in topic j.

5.2.5 Supervised Extractive Summarization Method (SESM)

The UESM's result on Section 5.3.1 shows that it produces generic summaries as intended. However, this study requires summaries focused on specific topics – namely (i.e., QMDS), trade-offs between provision and promotion strategies in CRCCP. Thus, UESM's current limitations lie in its inability to generate query-based or topic-specific summaries.

This section introduces the Supervised Extractive Summarization Method (SESM), an extractive summarization method that utilizes coded sentences from the content analysis (CA) process to guide topic-specific summary generation. We begin with a brief overview of the standard CA process, followed by the details of how SESM leverages these coded elements. Figure 5.1 illustrates the general CA workflow and how SESM incorporates coded sentences within this framework. Note that the CA team conducts the CA part of the method, which is shaded in grey.

Content analysis (CA)

Qualitative methodologies are used in healthcare to gain insights into complex phenomena and experiences of nurses, providers, policy-makers, and patients [25, 90, 104]. For example, a study from Jones et al. [53] employs qualitative methods to identify patients' challenges from their complex perspectives and experiences to increase the CRC screening uptake. These methodologies provide valuable information about the healthcare system from the perspective of those experiencing it [108]. One such methodology is content analysis (CA), which involves systematically analyzing large amounts of textual information through coding and categorization to identify patterns, frequency, relationships, and structure in the words used as well as the discourses of communication [68, 81]. In CA, coding refers to the process of systematically categorizing and analyzing text elements or word occurrence frequency to identify patterns, themes, or concepts within the text. Depending on the research aim, CA can be implemented with an inductive approach, where themes are identified from the data, or a deductive approach, which is useful for testing previous theories in different situations or comparing themes at different periods. [51].

Elo and Kyngäs [30] outline the CA process as encompassing three key stages: preparation, organization, and reporting, depicted in Figure 5.1. During the preparation phase, researchers familiarize themselves with the data, select the unit of analysis (e.g., words, sentences), and determine the focus – manifest (explicit content) or latent (underlying meaning). This initial stage involves identifying relevant textual data sources (documents, transcripts, etc.) aligned with the research objectives. These sources can be collected from archives, websites, interviews, or other locations. Researchers then decide on the level of detail to analyze, such as focusing on individual words or complete sentences. Additionally, they will choose whether to prioritize explicit content (manifest) or delve into latent content (underlying meaning or assumptions).

The organization phase involves open coding, category identification, grouping, and abstraction for a general description. Open coding entails assigning headings and notes while reviewing the material or highlighting sentences relevant to the research topic and representative of emerging themes. These coded labels can be interpreted as categories and grouped based on thematic similarities. Finally, the reporting phase involves presenting the findings through various methods, such as conceptual maps, category systems, or a storyline narrative.

While CA is a prevalent tool in healthcare research, it has limitations. These include overlooking textual context and potentially misinterpreting frequent themes as more significant [73, 108, 92]. To assist CA users in summarizing their coded sentences beyond theme frequency, we propose to employ the extractive summarization method detailed in Section 5.2.4 tailored for this task.

Coded sentences

In this research, our CA team employed CA with an inductive approach to answer the research question of identifying trade-offs of provision and promotion strategies based on the analyzed literature. The CA team coded the identified 38 relevant journal articles related to the Colorectal Cancer Control Program (CRCCP) published between 2011 and 2023, the same set of documents in UESM's exploration. Similar to the text data included in UESM from all sections in the paper, our CA team also comprehensively reviewed all sections within the papers. This thorough examination resulted in 399 coded sentences with an average of 41.43 words and a standard deviation of 21.64 words.

The CA process encompasses reading the papers, generating and consolidating themes between two analysts, and recording the results in a table. The initial familiarization with the papers required approximately one hour per document, while subsequent analysis averaged 30-45 minutes, depending on the paper length. This process was completed within 30 hours.

Each coded sentence can encompass one to three sentences from the original papers and can be assigned to one or more sub-themes. These 399 coded sentences were further categorized into 26 sub-themes, which were subsequently grouped into 6 overarching themes. They are as follows:

1. Recommendation for improving screening or testing qualities

2. Efficiency

3. Cost considerations

4. Selecting which EBIs to implement

5. Implementation and effectiveness of EBIs

6. Strategies for maximizing EBI duration or sustainability

Each trade-off theme relates to a different aspect in CRCCP (i.e., provision and promotion). "Recommendation for improving screening or testing qualities" particularly relates to the screening provision program since a well-monitored test ensures the completeness of the test provided. Moreover, "Selecting which EBIs to implement","Implementation and effectiveness of EBIs", and "Strategies for maximizing EBI duration or sustainability" are only related to promotion strategy because these trade-offs are related to EBIs. Lastly, "Efficiency" and "Cost consideration" relate to both screening provision and promotion because implementing either or both programs needs some resources and costs.

The Proposed Method with Adjusted Parameters

To assist CA users in summarizing the coded sentences, we propose leveraging the extractive summarization method detailed in Section 5.2.4: The Proposed Method, with an adjustment in the input text. This method analyzes the *coded sentences* as a whole unit where each coded sentence can encompass up to three sentences. Preserving their entirety allows us to capture the full context and meaning intended by the coder.

5.2.6 Evaluation metrics

The best way to evaluate a summary is by having a human review it. However, this approach is costly and time-consuming [62]. To address this issue, a popular metric called ROUGE (Recall-Oriented Understudy for Gisting Evaluation) by Lin [61] is often used to evaluate automatic text summarization [18]. ROUGE works by comparing the number of common words between a system-generated summary (candidate) and a human-made summary (reference). The reference summary is considered as the benchmark for comparison. The ROUGE metric is commonly used to evaluate the quality of summaries and has been found to have a strong correlation with human judgments [35]. However, it is limited as it only matches exact strings between the system-generated and human-made summaries without considering the meaning of words and synonyms [56].

Acknowledging the limitations of both human review and ROUGE scoring, we will employ a multifaceted evaluation approach using both methods to evaluate the final summaries generated by UESM and SESM. We evaluate the UESM through a human review since this summary will be used as a reference point for the SESM. In SESM, we will use human review and ROUGE score to evaluate the summary.

Human review

Three reviewers will evaluate the summary generated from the UESM and SESM. Two reviewers are from the CA team, and one reviewer is a PhD student in Health Policy and Management at the University of North Carolina who is familiar with CRCCP-related projects. The CA team has already reviewed all papers, providing them with the necessary context for this evaluation. Specifically, we will use four dimensions recommended by Kryściński et al. [57] to be considered for rating the generated summaries:

1. **Relevance** refers to the selection of important content from the source. The summary should include only important information from the source document. Reviewers were encouraged to penalize summaries that contained redundancies and excess information.

2. **Consistency** refers to the factual alignment between the summary and the source. A factually consistent summary contains only statements that are entailed by the source document. Reviewers were also asked to penalize summaries that contained hallucinated facts.

3. **Fluency** refers to the quality of individual sentences. It is expected that the summary does not have formatting problems, capitalization errors, or ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read.

4. **Coherence** refers to the collective quality of all sentences. The summary is expected to be well-structured and well-organized. It should not just be a heap of related information but should build from sentence to sentence to a coherent body of information about a topic.

In addition to the four dimensions evaluated, reviewers assessed the overall quality of the generated summaries. Specifically, they rated how well the summaries addressed the following question: "Does this paragraph summarize the trade-offs between provision and promotion strategies in CRCCP across the identified dimensions from the selected papers?" This assessment was applied to both UESM, which generates generic summaries, and SESM, which generates query-based summaries. All ratings, including those for the four dimensions and the overall review, used a 5-point Likert scale, with 1 indicating "bad" and 5 indicating "good."

The reviewers also provided individual reviews with comments to gain a deeper understanding of the selected coded sentences. In particular, we asked for their opinion on "What is lacking in the summary?" to provide a deeper understanding of the generated summary.

In a nutshell, we will provide a human evaluation that rates the relevance, consistency, fluency, coherence, and overall quality of the generated final summary on the Likert scale and comments on individual selected coded sentences.

Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

In addition to the human evaluation of the final summary generated by the SESM, we will also calculate the ROUGE metrics. As previously mentioned, the ROUGE metrics will calculate the level of word overlap between the candidate summary (i.e., the final summary generated by the SESM) and a reference summary (gold standard). This reference summary is created by concatenating the individual sub-theme summaries established by the CA team. Despite the ROUGE's limitation in detecting the words' semantic meaning between the candidate and reference summaries, we will calculate the ROUGE metrics as a reference for future study.

There are several ROUGE metrics, but in this study, we will focus on the ROUGE-1 and ROUGE-L metrics to assess the overlap.

1. **ROUGE-1 (R-1)**: This metric is based on the uni-gram measure, comparing candidate and reference summaries. ROUGE-N represents n-gram recall between a candidate summary and reference summaries.

2. **ROUGE-L (R-L)**: It relies on the longest common subsequences between a candidate and reference summaries.

In particular, we calculated the precision, recall, and F-1 scores for ROUGE-1 and ROUGE-L.

- **Precision** For R-1, precision is calculated as the ratio of the number of individual words or unigrams in the candidate summary that also appears in the reference summary over the number of unigrams in the *candidate* summary.

  For the R-L, we calculate the ratio of the longest common sub-sequences (LCS) in the candidate summary that also appears in the reference summary over the unigrams in the *candidate* summary.

  For example, if
  the reference (R) is "The cat is on the mat," and
  the candidate (C) is "The cat and the dog."

  Then, we calculate the precision for follow:

– For R-1, we divided the number of unigram appears in both R and C (“the”, “cat”, “the”) over the number of unigrams in C ("the", "cat", "and", "the", "dog"). So, the calculation will be

$$P = \frac{3}{5} = 0.6 \tag{5.7}$$

– For R-L, we divided the number of LCS appears in both R and C (“the cat the”) over the number of unigrams in C ("the", "cat", "and", "the", "dog"). So, the calculation will also be similar.

$$P = \frac{3}{5} = 0.6 \tag{5.8}$$

- **Recall** is calculated as the ratio of the number of unigrams or LCS in the reference summary that also appears in the candidate summary over the number of unigrams in the *reference* summary.

- **F1-score** is calculated directly from precision and recall score using the standard calculation. Each value ranges from 0 to 1, where 0 indicates no overlap at all, and 1 indicates perfect word overlap between the candidate and the reference summary.

5.3 Result

This study aims to develop a query-based summarization method. We employ a case study approach by applying the method to summarize 38 selected research papers on CRCCP to analyze the trade-offs between provision and promotion strategies in Colorectal Cancer Control Programs (CRCCPs). This section presents the results for unsupervised and supervised extractive summarization methods, referred to as UESM and SESM, respectively. We evaluate UESM with a human review and SESM with a human review and ROUGE scores.

5.3.1 Unsupervised Extractive Summarization Method (UESM)

Out of the 4616 sentences, we employed the UESM to select 13 sentences that were concatenated as the final summary. The entire summarization process, which encompasses preparation (e.g., text preprocessing, sentence embedding), and summarization steps, can be completed within 10 to 15 minutes.

Following is the 13-sentence summary generated by UESM for text in all sections of 38 selected papers.

"The application consisted of questions about FQHCs' current CRC screening processes and rates and the organizational capacity of the health systems to implement EBIs and partner with WVPICCS. Overall, grantees are implementing more client-oriented than provider-oriented strategies; efforts to help grantees implement provider-oriented strategies and reduce structural barriers to CRC screening may significantly increase screening rates. Our development process did not include clinic-level representatives, given the extraordinary bur- den clinics were facing because of the COVID-19 pandemic.CRC screening reduces cancer deaths by detecting cancers at an early stage and by detecting and removing precancerous polyps before cancer develops (2). Evaluation Methods CDC's Framework for Program Evaluation was applied to design the clinic survey on which this analysis is based (21).To guide these efforts, the American Society for Gastrointestinal Endoscopy/American College of Gastroenterology Task Force on Quality in Endoscopy recommended a subset of 3 high-priority indicators: 1) ADR in asymptomatic average-risk persons (screening), 2) frequency of colonoscopies following recommended surveillance and rescreening intervals, and 3) cecal intubation rate with photo documentation (5). Thirty funded awardees partnered with primary care clinics to implement evidence-based interventions (EBIs) recommended by the Community Preventive Services Task Force in The Community Guide, coupled with other supporting activities (SAs), to increase uptake of CRC screening (Centers for Disease Control and Prevention [CDC], 2016). Poverty rates (0.2) and percent of the Hispanic population percentage (0.1) in counties were also associated positively, while the percentage of non-Hispanic Black populations was associated negatively with clinic screening rates. Results from the GEE model showed a slight decrease in estimated coefficients and a slight increase in p-values. This difference may have been due to differences in eligible populations, interventions and their duration, methods and analytical approaches, and other factors. Direct clinical services-related activities's provision of screening tests, diagnostic services (diagnostic colonoscopy after positive FOBT or FIT), and surveillance procedures (follow up procedures after polyp or cancer diagnosis for individuals requiring surveillance); 2. We examined the total cost per person served by 3 cost components: total direct clinical cost, total direct nonclinical cost, and total indirect cost (12;14). The approach includes the collection of process measures, screening outcomes, cost and resource use data, and qualitative interviews to learn more about the design of programs and implementation procedures."

Furthermore, Table 5.1 shows the Likert scale results of the three reviewers for the generated summary from UESM. Out of the three reviewers, Reviewer 3 was not actively involved in the CA analysis. The table shows that for two reviewers, the relevance dimension has the lowest score (1 and 1.5), while fluency has the highest score (3). The low relevance score indicates that from the point of view of summarizing the trade-offs of the provision and promotion strategies, this summary does not select relevant sentences. On the other hand, the summary has good fluency, which indicates that the text is free from formatting problems that make it difficult to read. Moreover, the summary as a whole (i.e., overall) is rated 1.5 and 2 out of 5 by the reviewer. On the other hand, Reviewer 3 considered this summary to be good and gave an overall score of 4.5 with a maximum score for relevance and consistency (5) and second highest score for fluency and coherence (4).

Table 5.1: Table of Human Review with Likert Scale from 1 to 5 (Higher is better) for UESM

Dimensions	Reviewer 1	Reviewer 2	Reviewer 3
Relevance	1	1.5	5
Consistency	2	2	5
Fluency	3	3	4
Coherence	2	2	4
Overall	1.5	2	4.5

In addition to the evaluation of the summary as a whole, the following is the key insight from individually reviewing the selected sentences of the final summary by the human reviewer:

- **Lack of coherency**: The generated summary is considered to be confusing because the selected sentences describe specific ideas from different papers. Thus, the idea's flow is disorganized.
- **Lack of completeness**: While the summary mentioned the topic of costs and effectiveness, which are trade-offs identified by the CA team, further information is needed to understand the text.

5.3.2 Supervised Extractive Summarization Method (SESM)

As shown in Table 5.1, the relevance dimension of UESM's summaries received low ratings for 2 out of 3 reviewers. This suggests that the selected sentences are not relevant to summarizing the trade-offs between provision and promotion strategies in CRCCP. To address this limitation, we proposed using SESM.

In this method, we collected 399 coded sentences out of 38 selected papers from the CA process, reviewing all sections of the paper. Moreover, our research team (CA users) also created summaries for each subthemes as reference summaries or gold standards. Our CA process, including sentence coding and extractive summarization, can be completed within approximately 30 hours. Breaking down this time frame, sentence coding takes around 30 hours, while the extractive summarization itself is significantly faster, requiring only 10-15 minutes. In comparison, manual sub-theme summarization by the CA team took approximately 12-16 hours.

Following is the 13-coded sentence summary generated by SESM. Different from the previous summary from UESM, this summary has a variable length because each coded sentence can consist of up to three sentences. To distinguish these coded sentences, we use the symbol "||" for separation.

> "With improved tracking and automated reminder systems, mailed FIT kits paired with tailored patient education and clear instructions for completing the test may help primary care clinics catch up on the backlog of missed screenings during the COVID- 19 pandemic.|| Although CRC screening uptake at Neighborhood Healthcare was already at 75%, the interventions increased CRC screening uptake even more at a reasonable cost per additional person screened.|| The ultimate goal of using EBIs is to increase CRC screening; future research should examine the intensity with which these EBIs were implemented and to what extent CRC screening rates changed, especially in groups with low screening rates|| Overall, more grantees implemented client-oriented EBIs than provider-oriented EBIs (Table 2). By 2015 (program year 6), most grantees were implementing small media (n = 25) and client reminders (n = 26); few grantees stopped implementing these EBIs by 2015 (n = 4 for small media and n = 2 for client reminders).|| Study results support the need for awardees and implementation partners to provide health systems and clinics with tailored TA for their specific EHR system throughout the funding cycle to support EHR optimization for CRC screening, including avoiding common pitfalls (e.g., improper data entry, overly complex systems, changing EHRs) in ever-changing health care environments.|| By coordinating EBI and SA delivery with other disease prevention services, FQHCs more efficiently utilized resources; improved data reporting practices; minimized burden to FQHC staff and patients; and improved patient outcomes by increasing the likelihood of screening patients || The successful implementation of EBIs in primary care clinics is dependent on other activities such as quality improvement initiatives, effective electronic health systems, and good referral and patient tracking systems for colonoscopy.|| Of interest, interventions in-

volving all four EBIs or all strategies were not associated with the largest increases in CRC screening rates. Reasons for this could be diminishing returns of adding EBIs or a ceiling effect as clinics newly implementing or enhancing all EBIs or all strategies may have had higher baseline screening rates than those that newly implemented or enhanced fewer EBIs or SAs.|| Two aspects of the funding environment were identified as facilitating readiness for integrated implementation: (1) coordinating funding opportunities across multiple chronic disease programs to support consolidated application processes for clinic partners and (2) contracting with expert implementation partners to provide training and technical assistance to clinic partners that emphasized integrated implementation.|| Overall, FOBT/FIT-based programs and colonoscopy programs incurred substantial nonclinical costs per person served ($1,018 for colonoscopy and $980 for FIT/FOBT). Examples of non-clinical costs were managing contracts with providers and program management. These findings indicated that although the clinical cost of colonoscopy programs was higher than the clinical cost of FOBT/FIT programs, the cost of nonclinical services required to manage the programs and deliver the screenings was similar. || Second, the study does not account for cost per patient over an extended period to compare the long-term cost of colonoscopy versus FOBT/FIT-based programs. We only report cost for the first testing period (screening and diagnostic tests required), and our estimates do not provide the overall cost of FIT/FOBT and colonoscopy programs.|| Systematic reviews have identified barriers to CRC screening including low levels of education, language or communication issues, low socioeconomic status, lack of insurance coverage, and general attitudes towards prevention (for example, smokers are less likely to seek screening) (Gimeno Garcia, 2012; Subramanian et al., 2004). || Therefore, the screening promotion activities themselves accounted for about 65% of the total promotion costs and the remaining 35% was related to supporting programmatic activities such as program management and partnership development."

Furthermore, Table 5.2 shows the Likert scale results of the human review for the generated summary from SESM. The table shows that the coherence dimension has the lowest score (2 out of 5) for all reviewers. Moreover, fluency has the highest score (4 out of 5) for Reviewers 1 and 2, while relevance and consistency have a tie rank for Reviewer 3. Furthermore, the relevance score is higher than the generated summary from UESM for Reviewers 1 and 2 (with scores 1 and 1.5) but not for Reviewer 3. This indicates that from the point of view of summarizing the trade-offs of the provision and promotion strategies, the two reviewers consider this summary selected more relevant sentences than the UESM's generated summary. Moreover, the summary as a whole (i.e., overall) is rated 3 and 2.5 for Reviewers 1 and 2, which is higher than the overall rate for the summary generated by UESM. Yet, Reviewer 3 considered the SESM's summary (3) to be worse than the UESM's summary (4.5).

Table 5.2: Table of Human Review with Likert Scale from 1 to 5 (Higher is better) for SESM

Dimensions	Reviewer 1	Reviewer 2	Reviewer 3
Relevance	3	2.5	4
Consistency	4	3	4
Fluency	4	4	2
Coherence	2	2	2
Overall	3	2.5	3

Following are the key insights from individually reviewing the several coded sentences chosen from the SESM that are considered unsatisfactory:

1. **Missing some of the trade-offs**. While this summary is selected from coded sentences, some of the reviewers found that the summary missing "costs", "sustainability", and aspects of "EBI selection" trade-offs.

2. **A stand-alone coded sentences**. This type of coded sentence needs more context because the sentences only describe one piece of evidence.

3. **Lack of introductory sentences**. This point is mentioned by Reviewer 3, who was not actively involved in the CA effort.

In addition to the Likert scale and the comments from the reviewer, we calculated the ROUGE-1 and ROUGE-L metrics. These metrics compare a machine-generated summary (in this case, the method's output) to a human-generated summary considered the "gold standard." We concatenated all human-generated summaries for each subthemes into a single reference summary for evaluation. Similarly, the method's selected sentences were combined to form a single candidate summary. Table 5.3 presents the recall, precision, and F1 scores for ROUGE-1 and ROUGE-L.

Table 5.3: Table of ROUGE-1 and ROUGE-L score

ROUGE-1			ROUGE-L		
Recall	Precision	F-1	Recall	Precision	F-1
0.756	0.196	0.311	0.297	0.077	0.122

5.4 Discussion

This study proposed an extractive summarization (ES) method for query-based multi-documents (QMDS) text data called the supervised extractive summarization method (SESM). As a case study, we attempted to summarize the trade-offs between providing and promoting colorectal cancer (CRC) screening services from selected journal papers related to the topic. Popular abstractive methods, such as LLMs, can produce a well read summary by generating words. However, this method suffers from "hallucination" that is detrimental in the healthcare setting. Moreover, traditional content analysis (CA) can be used to summarize this topic from the literature, but the method is time-consuming and susceptible to missing contextual nuances. Therefore, the supervised extractive summarization method (SESM) is proposed to summarize the literature based on the user's query.

We first evaluated the UESM based on the work of Belwal, Rai, and Gupta [6]. We modified the sentence selection process to prioritize sentences across clusters and topics that maximize overall similarity scores. This approach aimed to minimize redundancy and enhance relevancy by covering various themes within the retrieved articles. However, two reviewers who were actively involved in CA consider the summary to be lacking in relevance (the relevance rate is 1 and 1.5 by two reviewers), failing to capture the specific topic of trade-offs between service provision and promotion strategies. Additionally, the lack of smooth transitions between sentences can hinder the overall coherence of the UESM summary (the coherence rate is 2 out of 5 by two reviewers and 4 out of 5 by one reviewer). Thus, the UESM's summary may not be ideal for capturing query-specific details (the overall rate is 1.5 and 2 out of 5 by two reviewers). On the other hand, one reviewer (i.e., Reviewer 3) favors this summary. This could be attributed to either a preference for the comprehensiveness of the CRCCP background information it provides or a general tendency to assign higher scores. Nevertheless, the tendency of producing a generic summary is understandable because the UESM generates the summary based on the topics covered across the sentences, including the introduction and unrelated sentences.

To address the limitation, we introduced SESM, which combines CA with ES. This method leverages the strengths of both approaches: ES assists CA in summarizing their coded sentences, while CA provides context for the summarization process. The SESM selected 13 sentences from a pool of 399 sentences to summarize the papers.

The UESM offers significant time savings (10 to 15 minutes) compared to CA (around 30 hours) and the SESM (30 hours). However, the UESM's summary lacks the necessary focus to summarize the trade-offs of the provision and promotion strategies of CRCCP. Therefore, the SESM offers benefits for content analysts seeking focused summaries.

The F-1 scores for ROUGE-1 (0.311) and ROUGE-L (0.122) indicated a reasonable overlap in individual words between the method-generated and human-generated summaries but not in the word sequence. In comparison, the F-1 of ROUGE-1 and ROUGE-L from the evaluation tests conducted by Belwal, Rai, and Gupta [6] are 0.431 and 0.395, which are higher than the SESM's score. The evaluation score from Belwal's study is based on summarizing a single document from CNN/DailyMail data [47] with their proposed extractive summarization. This data consists of information compiled from numerous news articles. However, as El-Kassas et al. [56] point out, these metrics don't account for the semantic meaning conveyed. Therefore, the ROUGE scores serve as a baseline for future studies employing more nuanced evaluation methods.

Human feedback suggested promising results for SESM, with the generated sentences being relevant to the identified sub-themes for two reviewers having been actively involved in the CA (relevance rate is 3, and 2.5 out of 5 compared to UESM with relevance rate of 1, and 1.5). On the other hand, one reviewer assigned a higher relevancy score (5) to the UESM summary compared to the SESM summary (4). This preference might be due to the UESM summary's inclusion of introductory sentences, which could provide context for readers unfamiliar with the CRCCP program. In contrast, the SESM summary, designed to be concise and focused on the user query, might lack those introductory elements. This difference could be particularly relevant for reviewers who were not actively involved in the content analysis (CA) process and might benefit from the additional background information provided by the UESM summary. Moreover, while the method utilizes a pre-selected pool of sentences from the CA stage, the challenge of summarizing coded sentences while considering context remains [73, 92, 108]. We believe this method can assist content analysts by providing a selection of relevant sentences to streamline the summarization process.

An interesting advantage of the proposed method is its user control over method parameters. For example, the number of clusters and topics generated by the method directly impacts the number of sentences included in the summary. However, defining a "good" summary length for CA users will be easier because the number of sentences in the final summary can be based on the number of themes or sub-themes as the lower bound. Additionally, the method allows users to trace topics within the summary, facilitating the organization phase of CA users when revisiting the themes and their groups might be necessary. This feature can prove valuable for situations where the reading process during the preparation phase might not reveal the most suitable thematic organization.

5.5 Limitations and Future Research

This study acknowledges three key limitations that future research can address. First, the pre-selection of the sentence pool from which the method draws its outputs grants it a certain advantage. All sentences within this pool have already been identified as relevant, potentially limiting the method's ability to demonstrate true selection capability. To address this, future work could develop a multi-label classification system for sentences. This system would assist content analysts in coding new sentences and provide the method with a more diverse and unbiased selection of sentences for summarization.

Second, the evaluation involved the generation of a single summary, precluding any comparative analysis with alternative methods or the method's performance on unseen data. Future research could address this limitation by employing a more comprehensive evaluation strategy. This strategy could involve comparing the proposed method's performance with other extractive summarization techniques or evaluating its ability to generate summaries for unseen datasets.

Third, exploring the impact of different sentence embedding or topic modeling techniques on the method's performance remains an interesting avenue for further investigation. Different embedding techniques capture various semantic relationships between words, while alternative topic modeling approaches might uncover latent thematic structures within the data. Evaluating the influence of these variations could lead to significant improvements in the method's effectiveness.

5.6 Conclusion

This study investigated the feasibility of a Supervised Extractive Summarization Method (SESM) for generating query-based summaries. We use a case study of capturing the trade-offs between service provision and promotion strategies in CRCCP programs as our query of interest. SESM leverages the strengths of both content analysis (CA) and extractive summarization (ES) to produce user-focused summaries. While the Un-supervised Extractive Summarization Method (UESM) offered significant time savings compared to CA, its summaries lacked the necessary focus on the user query. In contrast, SESM demonstrated promise in generating summaries relevant to user needs, as evidenced by human feedback.

However, limitations exist. The pre-selection of the sentence pool and the generation of a single summary for evaluation necessitate further investigation. Future research directions include developing a multi-label classification system for sentences, conducting a more comprehensive evaluation strategy, and exploring the impact of different sentence embedding or topic modeling techniques. Overall, SESM shows potential as a valuable tool for content analysts seeking efficient and focused summaries for CRCCP program analysis.

Overall, this study underscores the potential of extractive summarization techniques to streamline CA. Our proposed supervised method empowers content analysts by assisting them in summarizing large volumes of coded textual data, particularly when focused summaries capturing key themes are desired. While existing summarization tools can handle general documents, this study highlights the need for further exploration in developing methods tailored to specific topics within the text without "hallucination" as often happens in summaries generated by LLMs. By tackling this challenge, future research can unlock the full potential of extractive summarization for CA, and other qualitative research tasks.

CHAPTER

6

CONCLUSION

This dissertation explored the versatility of Natural Language Processing (NLP), specifically topic modeling, for analyzing various healthcare data sources beyond Electronic Health Records (EHRs). Due to privacy concerns and access limitations with EHRs, we investigated alternative resources: internal hospital communications during the COVID-19 pandemic, call transcripts from the diabetic retinopathy (DR) care coordination program, and journal articles on the Colorectal Cancer Control Program (CRCCP). We introduce the topic modeling method in Chapter 2, then continue elaborating on the exploration in other chapters.

In Chapter 3, we analyzed internal hospital communications during the early stages of COVID-19. Using structural topic modeling (STM), we identified twelve key topics, including visitor policies, best practice updates, and staff benefits. Interestingly, the two hospital systems differed in their communication focus, with both of them emphasizing operational adjustments. Our findings also revealed alignment with recommended communication topics from the American Hospital Association (AHA).

Chapter 4 investigated factors influencing enrollment in a care coordination program for DR patients. We analyzed call transcripts from the enrollment process using STM to identify program-related topics. We employed the penalized logistic regression model with the least absolute shrinkage and selection operator (Lasso) and cross-validated the model to predict the eventual enrollment of the call history. We discovered that discussions about scheduling counseling appointments had high odds for enrollment, while conversations focused solely on greetings and program explanations had low odds. This suggests the importance of clear action steps for patients considering the program.

Lastly, in Chapter 5, we attempted to address the limitations of abstractive summarization methods, particularly "hallucinations", which can be misleading in healthcare settings. We introduced a Supervised Extractive Summarization Method (SESM) that combines the strengths of Content Analysis (CA) and extractive summarization to create a query-based multi-document summary (QMDS). As a baseline, we introduce the Unsupervised Extractive Summarization Method (UESM), which generates a generic summary adapted from a previous study. Our case study attempted to summarize trade-offs between the provision and promotion strategies of Colorectal Cancer Control Programs (CRCCP) from 37 academic papers. We compared UESM with SESM to demonstrate SESM's ability to generate user-query-specific summaries. Human evaluation indicated promising results for SESM's relevance to user needs. While limitations exist, such as using pre-selected sentence pools from CA and the need for testing the method on different datasets, this study highlights the potential of extractive summarization, particularly the supervised model (SESM), for focused summarization in healthcare text analysis. Future research can explore multi-label classification for sentence coding and investigate the impact of different sentence embedding or topic modeling techniques on SESM's performance.

Overall, this dissertation demonstrates the potential of NLP, particularly topic modeling, for analyzing diverse healthcare data sources. Our studies showcased the ability of NLP to uncover communication patterns during a pandemic, identify factors influencing patient behavior in a care coordination program, and streamline content analysis for focused summarization of scientific literature. By overcoming challenges associated with EHR access, this research paves the way for further exploration of NLP applications in healthcare data analysis, ultimately aiming to improve patient care and program effectiveness.

BIBLIOGRAPHY

[1] Centers for Disease Control and Prevention (CDC). *U.S. Cancer Statistics: Colorectal Cancer Stat Bite | CDC.* en-us. June 2023. URL: https://www.cdc.gov/cancer/uscs/about/stat-bites/stat-bite-colorectal.htm (visited on 11/21/2023).

[2] U.S. Department of Health and Human Services (USDHHS). *Increase the proportion of adults who get screened for colorectal cancer — C-07 - Healthy People 2030 | health.gov.* 2021. URL: https://health.gov/healthypeople/objectives-and-data/browse-objectives/cancer/increase-proportion-adults-who-get-screened-colorectal-cancer-c-07 (visited on 11/20/2023).

[3] American Hospital Association AHA. *Communications: Internal and External.* 2020. URL: https://www.aha.org/system/files/media/file/2020/07/aha-covid19-pathways-comms-internal-external.pdf.

[4] Edoardo M. Airoldi and Jonathan M. Bischof. "Improving and Evaluating Topic Models and Other Models of Text". en. In: *Journal of the American Statistical Association* 111.516 (Oct. 2016), pp. 1381–1403. ISSN: 0162-1459, 1537-274X. DOI: 10.1080/01621459.2015.1051182. URL: https://www.tandfonline.com/doi/full/10.1080/01621459.2015.1051182 (visited on 09/21/2023).

[5] Entesar Alanzi and Safa Alballaa. "Query-Focused Multi-document Summarization Survey". en. In: *International Journal of Advanced Computer Science and Applications* 14.6 (2023). ISSN: 21565570, 2158107X. DOI: 10.14569/IJACSA.2023.0140688. URL: http://thesai.org/Publications/ViewPaper?Volume=14&Issue=6&Code=IJACSA&SerialNo=88 (visited on 06/24/2024).

[6] Ramesh Chandra Belwal, Sawan Rai, and Atul Gupta. "Extractive text summarization using clustering-based topic modeling". en. In: *Soft Computing* 27.7 (Apr. 2023), pp. 3965–3982. ISSN: 1433-7479. DOI: 10.1007/s00500-022-07534-6. URL: https://doi.org/10.1007/s00500-022-07534-6 (visited on 11/21/2023).

[7] Kenneth Benoit et al. *quanteda: Quantitative Analysis of Textual Data.* May 2023. URL: https://cloud.r-project.org/web/packages/quanteda/index.html (visited on 11/15/2023).

[8] BigScience Workshop. "BLOOM". In: (2022). DOI: 10.57967/HF/0003. URL: https://huggingface.co/bigscience/bloom (visited on 06/25/2024).

[9] Pratik K. Biswas and Aleksandr Iakubovich. "Extractive Summarization of Call Transcripts". In: *IEEE Access* 10 (2022), pp. 119826–119840. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2022.3221404.

[10] David M. Blei and John D. Lafferty. "A correlated topic model of Science". In: *The Annals of Applied Statistics* 1.1 (June 2007), pp. 17–35. ISSN: 1932-6157, 1941-7330. DOI: 10.1214/07-AOAS114. URL: https://projecteuclid.org/journals/annals-of-applied-statistics/volume-1/issue-1/A-correlated-topic-model-of-Science/10.1214/07-AOAS114.full (visited on 09/21/2023).

[11] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation". In: *The Journal of Machine Learning Research* 3.null (Mar. 2003), pp. 993–1022. ISSN: 1532-4435.

[12] Elliot Bolton et al. *BioMedLM: a Domain Specific Large Language Model for Biomedical Text. Stan ford CRFM Blog 2022*. URL: https://crfm.stanford.edu/2022/12/15/biomedlm.html (visited on 06/25/2024).

[13] Cristian Buciluă, Rich Caruana, and Alexandru Niculescu-Mizil. "Model compression". In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD '06. New York, NY, USA: Association for Computing Machinery, Aug. 2006, pp. 535–541. ISBN: 9781595933393. DOI: 10.1145/1150402.1150464. URL: https://doi.org/10.1145/1150402.1150464 (visited on 04/08/2024).

[14] CDC. *Diabetes and vision loss. Centers for Disease Control and Prevention*. Dec. 2022. URL: https://www.cdc.gov/diabetes/managing/diabetes-vision-loss.html (visited on 09/21/2023).

[15] Division of Cancer Prevention CDC. *Colorectal Cancer Control Program (CRCCP) | CDC*. en-us. Mar. 2023. URL: https://www.cdc.gov/cancer/crccp/index.htm (visited on 11/21/2023).

[16] Division of Cancer Prevention CDC. *Evidence-Based Interventions | CDC*. en-us. Dec. 2022. URL: https://www.cdc.gov/cancer/community-resources/interventions/index.htm (visited on 11/21/2023).

[17] CDC, Division of Cancer Prevention. *What Should I Know About Screening for Colorectal Cancer?* en-us. Feb. 2023. URL: https://www.cdc.gov/cancer/colorectal/basic_info/screening/index.htm (visited on 11/20/2023).

[18] Andrea Chaves, Cyrille Kesiku, and Begonya Garcia-Zapirain. "Automatic Text Summarization of Biomedical Text Data: A Systematic Review". en. In: *Information* 13.8 (Aug. 2022), p. 393. ISSN: 2078-2489. DOI: 10.3390/info13080393. URL: https://www.mdpi.com/2078-2489/13/8/393 (visited on 11/21/2023).

[19] Rob Churchill and Lisa Singh. “The Evolution of Topic Modeling”. en. In: *ACM Computing Surveys* 54.10s (Jan. 2022), pp. 1–35. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3507900. URL: https://dl.acm.org/doi/10.1145/3507900 (visited on 01/17/2024).

[20] Mark A. Clapp et al. “Natural language processing of admission notes to predict severe maternal morbidity during the delivery encounter”. In: *American Journal of Obstetrics and Gynecology* 227.3 (Sept. 2022), 511.e1–511.e8. ISSN: 0002-9378. DOI: 10.1016/j.ajog.2022.04.008. URL: https://www.sciencedirect.com/science/article/pii/S0002937822002800 (visited on 09/21/2023).

[21] Arman Cohan et al. “A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents”. In: (2018). DOI: 10.48550/ARXIV.1804.05685. URL: https://arxiv.org/abs/1804.05685 (visited on 06/24/2024).

[22] Guide to Community Preventive Services. *Summary CPSTF Findings Table for Cancer Prevention and Control | The Community Guide*. en-us. Jan. 2023. URL: https://www.thecommunityguide.org/pages/task-force-findings-cancer-prevention-and-control.html (visited on 11/21/2023).

[23] Natalie M. Davoodi, Margaret Healy, and Elizabeth M. Goldberg. “Rural America’s Hospitals are Not Prepared to Protect Older Adults From a Surge in COVID-19 Cases”. en. In: *Gerontology and Geriatric Medicine* 6 (Jan. 2020), p. 233372142093616. ISSN: 2333-7214, 2333-7214. DOI: 10.1177/2333721420936168. URL: http://journals.sagepub.com/doi/10.1177/2333721420936168 (visited on 09/24/2023).

[24] Janez Demšar. “Statistical Comparisons of Classifiers over Multiple Data Sets”. In: *Journal of Machine Learning Research* 7.1 (2006), pp. 1–30. ISSN: 1533-7928. URL: http://jmlr.org/papers/v7/demsar06a.html (visited on 07/08/2024).

[25] Norman K. Denzin and Yvonna S. Lincoln. *The SAGE Handbook of Qualitative Research*. en. Google-Books-ID: AIRpMHgBYqIC. SAGE, Apr. 2011. ISBN: 9781412974172.

[26] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv:1810.04805 [cs]. May 2019. URL: http://arxiv.org/abs/1810.04805 (visited on 04/09/2024).

[27] Gemma Doleman et al. “Staff perceptions of the effectiveness of managerial communication during the COVID-19 pandemic: A cross-sectional study”. en. In: *International Journal of Nursing Practice* 29.4 (Aug. 2023), e13149. ISSN: 1322-7114, 1440-172X. DOI: 10.1111/ijn.13149. URL: https://onlinelibrary.wiley.com/doi/10.1111/ijn.13149 (visited on 09/23/2023).

[28] Noman Dormosh et al. "Predicting future falls in older people using natural language processing of general practitioners' clinical notes". eng. In: *Age and Ageing* 52.4 (Apr. 2023), afad046. ISSN: 1468-2834. DOI: 10.1093/ageing/afad046.

[29] Elizabeth G. Eakin et al. "Living Well With Diabetes: 24-Month Outcomes From a Randomized Trial of Telephone-Delivered Weight Loss and Physical Activity Intervention to Improve Glycemic Control". en. In: *Diabetes Care* 37.8 (Aug. 2014), pp. 2177–2185. ISSN: 0149-5992, 1935-5548. DOI: 10.2337/dc13-2427. URL: https://diabetesjournals.org/care/article/37/8/2177/29808/Living-Well-With-Diabetes-24-Month-Outcomes-From-a (visited on 09/21/2023).

[30] Satu Elo and Helvi Kyngäs. "The qualitative content analysis process". en. In: *Journal of Advanced Nursing* 62.1 (Apr. 2008), pp. 107–115. ISSN: 0309-2402, 1365-2648. DOI: 10.1111/j.1365-2648.2007.04569.x. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2648.2007.04569.x (visited on 04/07/2024).

[31] Paul Fairie et al. "Categorising patient concerns using natural language processing techniques". In: *BMJ Health & Care Informatics* 28.1 (June 2021), e100274. ISSN: 2632-1009. DOI: 10.1136/bmjhci-2020-100274. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8246286/ (visited on 09/23/2023).

[32] Joshua C. Feblowitz et al. "Summarization of clinical information: a conceptual model". eng. In: *Journal of Biomedical Informatics* 44.4 (Aug. 2011), pp. 688–699. ISSN: 1532-0480. DOI: 10.1016/j.jbi.2011.03.008.

[33] Daniel J. Feller et al. "Using Clinical Notes and Natural Language Processing for Automated HIV Risk Assessment". In: *Journal of acquired immune deficiency syndromes (1999)* 77.2 (Feb. 2018), pp. 160–166. ISSN: 1525-4135. DOI: 10.1097/QAI.0000000000001580. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5762388/ (visited on 09/21/2023).

[34] Jerome Friedman et al. *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models*. Aug. 2023. URL: https://cloud.r-project.org/web/packages/glmnet/index.html (visited on 11/15/2023).

[35] Mahak Gambhir and Vishal Gupta. "Recent automatic text summarization techniques: a survey". en. In: *Artificial Intelligence Review* 47.1 (Jan. 2017), pp. 1–66. ISSN: 1573-7462. DOI: 10.1007/s10462-016-9475-9. URL: https://doi.org/10.1007/s10462-016-9475-9 (visited on 11/25/2023).

[36] Yanjun Gao et al. “A scoping review of publicly available language tasks in clinical natural language processing”. In: *Journal of the American Medical Informatics Association : JAMIA* 29.10 (Aug. 2022), pp. 1797–1806. ISSN: 1067-5027. DOI: 10.1093/jamia/ocac127. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9471718/ (visited on 11/14/2023).

[37] Edward I. George and Robert E. McCulloch. “Variable Selection via Gibbs Sampling”. en. In: *Journal of the American Statistical Association* 88.423 (Sept. 1993), pp. 881–889. ISSN: 0162-1459, 1537-274X. DOI: 10.1080/01621459.1993.10476353. URL: http://www.tandfonline.com/doi/abs/10.1080/01621459.1993.10476353 (visited on 04/17/2024).

[38] Marzyeh Ghassemi et al. “Unfolding physiological state: mortality modelling in intensive care units”. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD ’14. New York, NY, USA: Association for Computing Machinery, Aug. 2014, pp. 75–84. ISBN: 9781450329569. DOI: 10.1145/2623330.2623742. URL: https://doi.org/10.1145/2623330.2623742 (visited on 09/21/2023).

[39] Ana D. Goode, Marina M. Reeves, and Elizabeth G. Eakin. “Telephone-Delivered Interventions for Physical Activity and Dietary Behavior Change: An Updated Systematic Review”. In: *American Journal of Preventive Medicine* 42.1 (Jan. 2012), pp. 81–88. ISSN: 0749-3797. DOI: 10.1016/j.amepre.2011.08.025. URL: https://www.sciencedirect.com/science/article/pii/S0749379711007422 (visited on 09/21/2023).

[40] Google. *Gemini - chat to supercharge your ideas*. en. URL: https://gemini.google.com (visited on 06/24/2024).

[41] Nihit Goyal and Michael Howlett. ““Measuring the Mix” of Policy Responses to COVID-19: Comparative Policy Analysis Using Topic Modelling”. en. In: *Journal of Comparative Policy Analysis: Research and Practice* 23.2 (Mar. 2021), pp. 250–261. ISSN: 1387-6988, 1572-5448. DOI: 10.1080/13876988.2021.1880872. URL: https://www.tandfonline.com/doi/full/10.1080/13876988.2021.1880872 (visited on 11/14/2023).

[42] Nicholas Graves et al. “Cost-Effectiveness of a Telephone-Delivered Intervention for Physical Activity and Diet”. en. In: *PLoS ONE* 4.9 (Sept. 2009). Ed. by German Malaga, e7135. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0007135. URL: https://dx.plos.org/10.1371/journal.pone.0007135 (visited on 09/21/2023).

[43] U.S. Cancer Statistics Working Group. *USCS Data Visualizations*. en. Nov. 2023. URL: https://www.cdc.gov/cancer/dataviz (visited on 11/20/2023).

[44] Christina Haag et al. "Blending citizen science with natural language processing and machine learning: Understanding the experience of living with multiple sclerosis". en. In: *PLOS Digital Health* 2.8 (Aug. 2023), e0000305. ISSN: 2767-3170. DOI: 10.1371/journal.pdig.0000305. URL: https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000305 (visited on 09/21/2023).

[45] Martin Härter et al. "Effectiveness of Telephone-Based Health Coaching for Patients with Chronic Conditions: A Randomised Controlled Trial". en. In: *PLOS ONE* 11.9 (Sept. 2016), e0161269. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0161269. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0161269 (visited on 09/21/2023).

[46] Abby Haynes et al. "Using self-determination theory to understand and improve recruitment for the Coaching for Healthy Ageing (CHAnGE) trial". en. In: *PLOS ONE* 16.11 (Nov. 2021), e0259873. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0259873. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0259873 (visited on 09/21/2023).

[47] Karl Moritz Hermann et al. "Teaching Machines to Read and Comprehend". In: (2015). DOI: 10.48550/ARXIV.1506.03340. URL: https://arxiv.org/abs/1506.03340 (visited on 04/17/2024).

[48] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. *Distilling the Knowledge in a Neural Network*. arXiv:1503.02531 [cs, stat]. Mar. 2015. DOI: 10.48550/arXiv.1503.02531. URL: http://arxiv.org/abs/1503.02531 (visited on 04/09/2024).

[49] Samuel Hoffstaetter. *pytesseract: Python-tesseract is a python wrapper for Google's Tesseract-OCR*. URL: https://github.com/madmaze/pytesseract (visited on 11/19/2023).

[50] Kimia Honarmand et al. "Personal, professional, and psychological impact of the COVID-19 pandemic on hospital workers: A cross-sectional survey". en. In: *PLOS ONE* 17.2 (Feb. 2022), e0263438. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0263438. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0263438 (visited on 09/23/2023).

[51] Hsiu-Fang Hsieh and Sarah E. Shannon. "Three approaches to qualitative content analysis". eng. In: *Qualitative Health Research* 15.9 (Nov. 2005), pp. 1277–1288. ISSN: 1049-7323. DOI: 10.1177/1049732305276687.

[52] Jinpeng Hu et al. "Improving Radiology Summarization with Radiograph and Anatomy Prompts". In: (2022). DOI: 10.48550/ARXIV.2210.08303. URL: https://arxiv.org/abs/2210.08303 (visited on 06/24/2024).

[53] Resa M. Jones et al. "Patient-Reported Barriers to Colorectal Cancer Screening". en. In: *American Journal of Preventive Medicine* 38.5 (May 2010), pp. 508–516. ISSN: 07493797. DOI: 10.1016/j.amepre.2010.01.021. URL: https://linkinghub.elsevier.com/retrieve/pii/S074937971000098X (visited on 04/12/2024).

[54] Sarah R. Jordan et al. "Forgotten frontline workers: Environmental health service employees' perspectives on working during the COVID-19 pandemic". en. In: *Journal of Hospital Medicine* 17.3 (Mar. 2022), pp. 158–168. ISSN: 1553-5592, 1553-5606. DOI: 10.1002/jhm.12781. URL: https://shmpublications.onlinelibrary.wiley.com/doi/10.1002/jhm.12781 (visited on 09/23/2023).

[55] EunKyo Kang, Narae Song, and HyoRim Ju. "Contents and sentiment analysis of newspaper articles and comments on telemedicine in Korea: Before and after of COVID-19 outbreak". en. In: *Health Informatics Journal* 28.1 (Jan. 2022), p. 146045822210755. ISSN: 1460-4582, 1741-2811. DOI: 10.1177/14604582221075549. URL: http://journals.sagepub.com/doi/10.1177/14604582221075549 (visited on 09/23/2023).

[56] Wafaa S. El-Kassas et al. "Automatic text summarization: A comprehensive survey". In: *Expert Systems with Applications* 165 (Mar. 2021), p. 113679. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2020.113679. URL: https://www.sciencedirect.com/science/article/pii/S0957417420305030 (visited on 11/21/2023).

[57] Wojciech Kryściński et al. "Neural Text Summarization: A Critical Evaluation". In: (2019). DOI: 10.48550/ARXIV.1908.08960. URL: https://arxiv.org/abs/1908.08960 (visited on 04/16/2024).

[58] M Kursite et al. "Internal communication within the healthcare system during the COVID-19 pandemic in Latvia". In: *European Journal of Public Health* 31.Supplement_3 (Oct. 2021), ckab164.582. ISSN: 1101-1262, 1464-360X. DOI: 10.1093/eurpub/ckab164.582. URL: https://academic.oup.com/eurpub/article/doi/10.1093/eurpub/ckab164.582/6405179 (visited on 09/23/2023).

[59] Yue Tong Leung and Farzad Khalvati. "Exploring COVID-19–Related Stressors: Topic Modeling Study". EN. In: *Journal of Medical Internet Research* 24.7 (July 2022), e37142. DOI: 10.2196/37142. URL: https://www.jmir.org/2022/7/e37142 (visited on 09/23/2023).

[60] Jianning Li et al. “ChatGPT in healthcare: A taxonomy and systematic review”. eng. In: *Computer Methods and Programs in Biomedicine* 245 (Mar. 2024), p. 108013. ISSN: 1872-7565. DOI: 10.1016/j.cmpb.2024.108013.

[61] Chin-Yew Lin. “ROUGE: A Package for Automatic Evaluation of Summaries”. In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013 (visited on 03/19/2024).

[62] Elena Lloret, Laura Plaza, and Ahmet Aker. “The challenging task of summary evaluation: an overview”. en. In: *Language Resources and Evaluation* 52.1 (Mar. 2018), pp. 101–148. ISSN: 1574-0218. DOI: 10.1007/s10579-017-9399-2. URL: https://doi.org/10.1007/s10579-017-9399-2 (visited on 11/25/2023).

[63] Saskia Locke et al. “Natural language processing in medicine: A review”. In: *Trends in Anaesthesia and Critical Care* 38 (June 2021), pp. 4–9. ISSN: 2210-8440. DOI: 10.1016/j.tacc.2021.02.007. URL: https://www.sciencedirect.com/science/article/pii/S2210844021000411 (visited on 11/14/2023).

[64] Renqian Luo et al. “BioGPT: generative pre-trained transformer for biomedical text generation and mining”. en. In: *Briefings in Bioinformatics* 23.6 (Nov. 2022), bbac409. ISSN: 1467-5463, 1477-4054. DOI: 10.1093/bib/bbac409. URL: https://academic.oup.com/bib/article/doi/10.1093/bib/bbac409/6713511 (visited on 06/25/2024).

[65] Chong Ma et al. “An Iterative Optimizing Framework for Radiology Report Summarization with ChatGPT”. In: *IEEE Transactions on Artificial Intelligence* (2024). arXiv:2304.08448 [cs], pp. 1–12. ISSN: 2691-4581. DOI: 10.1109/TAI.2024.3364586. URL: http://arxiv.org/abs/2304.08448 (visited on 06/24/2024).

[66] Congbo Ma et al. “Multi-document Summarization via Deep Learning Techniques: A Survey”. In: *ACM Computing Surveys* 55.5 (Dec. 2022), 102:1–102:37. ISSN: 0360-0300. DOI: 10.1145/3529754. URL: https://doi.org/10.1145/3529754 (visited on 04/06/2024).

[67] Annette E. Maxwell et al. “Promotion and Provision of Colorectal Cancer Screening: A Comparison of Colorectal Cancer Control Program Grantees and Nongrantees, 2011–2012”. en. In: *Preventing Chronic Disease* 11.10 (2014), e170. ISSN: 1545-1151. DOI: 10.5888/pcd11.140183. URL: https://escholarship.org/uc/item/3rk0z5hw (visited on 11/21/2023).

[68] Philipp Mayring. “Qualitative Content Analysis”. en. In: *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* Vol 1 (June 2000), No 2 (2000): Qualitative Methods in Various Disciplines I: Psychology. DOI: 10.17169/FQS-1.2.1089. URL: http://www.qualitative-research.net/index.php/fqs/article/view/1089 (visited on 04/07/2024).

[69] Jesse McLean et al. “Pandemic Responsiveness in an Acute Care Setting: A Community Hospital’s Utilization of Operational Resources During COVID-19”. In: *Journal of Multidisciplinary Healthcare* 15 (June 2022), pp. 1309–1321. ISSN: 1178-2390. DOI: 10.2147/JMDH.S361896. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9206458/ (visited on 02/07/2024).

[70] David Mimno et al. “Optimizing semantic coherence in topic models”. In: *Proceedings of the conference on empirical methods in natural language processing*. EMNLP ’11. Place: USA. Association for Computational Linguistics, July 2011, pp. 262–272. ISBN: 978-1-937284-11-4. (Visited on 09/23/2023).

[71] Rashmi Mishra et al. “Text summarization in the biomedical domain: A systematic review of recent research”. en. In: *Journal of Biomedical Informatics* 52 (Dec. 2014), pp. 457–467. ISSN: 15320464. DOI: 10.1016/j.jbi.2014.06.009. URL: https://linkinghub.elsevier.com/retrieve/pii/S1532046414001476 (visited on 06/13/2024).

[72] Diego Molla, Christopher Jones, and Vincent Nguyen. “Query Focused Multi-document Summarisation of Biomedical Texts”. In: (2020). DOI: 10.48550/ARXIV.2008.11986. URL: https://arxiv.org/abs/2008.11986 (visited on 06/26/2024).

[73] David L. Morgan. “Qualitative Content Analysis: A Guide to Paths not Taken”. en. In: *Qualitative Health Research* 3.1 (Feb. 1993), pp. 112–121. ISSN: 1049-7323, 1552-7557. DOI: 10.1177/104973239300300107. URL: http://journals.sagepub.com/doi/10.1177/104973239300300107 (visited on 04/07/2024).

[74] James Munkres. “Algorithms for the Assignment and Transportation Problems”. In: *Journal of the Society for Industrial and Applied Mathematics* 5.1 (1957), pp. 32–38. ISSN: 0368-4245. URL: https://www.jstor.org/stable/2098689 (visited on 04/14/2024).

[75] OASH. *Increase the proportion of adults with diabetes who have a yearly eye exam — D-04 - Healthy People 2030 | health.gov*. *Healthy People 2030*. URL: https://health.gov/healthypeople/objectives-and-data/browse-objectives/diabetes/increase-proportion-adults-diabetes-who-have-yearly-eye-exam-d-04 (visited on 09/21/2023).

[76] OpenAI et al. *GPT-4 Technical Report*. arXiv:2303.08774 [cs]. Mar. 2024. URL: http://arxiv.org/abs/2303.08774 (visited on 06/24/2024).

[77] Lai Meng Ow Yong et al. "Perception survey of crisis and emergency risk communication in an acute hospital in the management of COVID-19 pandemic in Singapore". In: *BMC Public Health* 20.1 (Dec. 2020), p. 1919. ISSN: 1471-2458. DOI: 10.1186/s12889-020-10047-2. URL: https://doi.org/10.1186/s12889-020-10047-2 (visited on 09/23/2023).

[78] Bo Pang et al. "Long Document Summarization with Top-down and Bottom-up Inference". In: (2022). DOI: 10.48550/ARXIV.2203.07586. URL: https://arxiv.org/abs/2203.07586 (visited on 06/24/2024).

[79] Rimma Pivovarov and Noémie Elhadad. "Automated methods for the summarization of electronic health records". eng. In: *Journal of the American Medical Informatics Association: JAMIA* 22.5 (Sept. 2015), pp. 938–947. ISSN: 1527-974X. DOI: 10.1093/jamia/ocv032.

[80] Joaquín Pizarro, Elisa Guerrero, and Pedro L. Galindo. "Multiple comparison procedures applied to model selection". en. In: *Neurocomputing* 48.1-4 (Oct. 2002), pp. 155–173. ISSN: 09252312. DOI: 10.1016/S0925-2312(01)00653-1. URL: https://linkinghub.elsevier.com/retrieve/pii/S0925231201006531 (visited on 07/08/2024).

[81] Catherine Pope, Sue Ziebland, and Nicholas Mays. "Analysing qualitative data". In: *BMJ : British Medical Journal* 320.7227 (Jan. 2000), pp. 114–116. ISSN: 0959-8138. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1117368/ (visited on 04/07/2024).

[82] Nils Reimers and Iryna Gurevych. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. arXiv:1908.10084 [cs]. Aug. 2019. DOI: 10.48550/arXiv.1908.10084. URL: http://arxiv.org/abs/1908.10084 (visited on 04/14/2024).

[83] Emil Rijcken et al. "Topic Modeling for Interpretable Text Classification From EHRs". In: *Frontiers in Big Data* 5 (2022). ISSN: 2624-909X. URL: https://www.frontiersin.org/articles/10.3389/fdata.2022.846930 (visited on 09/21/2023).

[84] Margaret Roberts et al. *stm: Estimation of the Structural Topic Model*. Aug. 2023. URL: https://cloud.r-project.org/web/packages/stm/index.html (visited on 11/15/2023).

[85] Margaret E. Roberts, Brandon M. Stewart, and Edoardo M. Airoldi. "A model of text for experimentation in the social sciences". In: *Journal of the American Statistical Association* 111.515 (July 2016), pp. 988–1003. ISSN: 0162-1459, 1537-274X. DOI: 10.1080/

01621459.2016.1141684. URL: https://www.tandfonline.com/doi/full/10.1080/01621459.2016.1141684 (visited on 09/21/2023).

[86] Prasenjeet Roy and Suman Kundu. "Review on Query-focused Multi-document Summarization (QMDS) with Comparative Analysis". en. In: *ACM Computing Surveys* 56.1 (Jan. 2024), pp. 1–38. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3597299. URL: https://dl.acm.org/doi/10.1145/3597299 (visited on 06/24/2024).

[87] Susan A. Sabatino et al. "Cancer Screening Test Receipt — United States, 2018". In: *MMWR. Morbidity and Mortality Weekly Report* 70.2 (Jan. 2021), pp. 29–35. ISSN: 0149-2195, 1545-861X. DOI: 10.15585/mmwr.mm7002a1. URL: http://www.cdc.gov/mmwr/volumes/70/wr/mm7002a1.htm?s_cid=mm7002a1_w (visited on 11/20/2023).

[88] Naoki Sakane et al. "Effects of telephone-delivered lifestyle support on the development of diabetes in participants at high risk of type 2 diabetes: J-DOIT1, a pragmatic cluster randomised trial". en. In: *BMJ Open* 5.8 (Aug. 2015), e007316. ISSN: 2044-6055, 2044-6055. DOI: 10.1136/bmjopen-2014-007316. URL: https://bmjopen.bmj.com/content/5/8/e007316 (visited on 09/21/2023).

[89] Malik Sallam. "ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns". eng. In: *Healthcare (Basel, Switzerland)* 11.6 (Mar. 2023), p. 887. ISSN: 2227-9032. DOI: 10.3390/healthcare11060887.

[90] Margarete Sandelowski and Julie Barroso. "Writing the proposal for a qualitative research methodology project". eng. In: *Qualitative Health Research* 13.6 (July 2003), pp. 781–820. ISSN: 1049-7323. DOI: 10.1177/1049732303013006003.

[91] Krishna P. Sharma et al. "Evidence-Based Interventions and Colorectal Cancer Screening Rates: The Colorectal Cancer Screening Program, 2015–2017". In: *American Journal of Preventive Medicine* 61.3 (Sept. 2021), pp. 402–409. ISSN: 0749-3797. DOI: 10.1016/j.amepre.2021.03.002. URL: https://www.sciencedirect.com/science/article/pii/S0749379721001938 (visited on 11/21/2023).

[92] Linda Shields and Alison Twycross. "Content analysis." English. In: *Paediatric Nursing* 20.6 (July 2008), pp. 38–39. ISSN: 09629513. URL: https://go.gale.com/ps/i.do?p=AONE&sw=w&issn=09629513&v=2.1&it=r&id=GALE%7CA181072637&sid=googleScholar&linkaccess=abs (visited on 04/07/2024).

[93] Hoo-Chang Shin et al. “BioMegatron: Larger Biomedical Domain Language Model”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber et al. Online: Association for Computational Linguistics, Nov. 2020, pp. 4700–4706. DOI: 10.18653/v1/2020.emnlp-main.379. URL: https://aclanthology.org/2020.emnlp-main.379 (visited on 06/25/2024).

[94] Natasha Smallwood et al. “Moral Distress and Perceived Community Views Are Associated with Mental Health Symptoms in Frontline Health Workers during the COVID-19 Pandemic”. en. In: *International Journal of Environmental Research and Public Health* 18.16 (Jan. 2021), p. 8723. ISSN: 1660-4601. DOI: 10.3390/ijerph18168723. URL: https://www.mdpi.com/1660-4601/18/16/8723 (visited on 09/23/2023).

[95] Akshay Smit et al. *CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT*. arXiv:2004.09167 [cs]. Oct. 2020. URL: http://arxiv.org/abs/2004.09167 (visited on 06/24/2024).

[96] Rachel L. Snyder et al. “Burnout and staff turnover among certified nursing assistants working in acute care hospitals during the COVID-19 pandemic”. en. In: *PLOS ONE* 18.8 (2023). DOI: 10.1371/journal.pone.0290880. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10468035/ (visited on 02/07/2024).

[97] Sujha Subramanian et al. “Costs of colorectal cancer screening provision in CDC’s Colorectal Cancer Control Program: Comparisons of colonoscopy and FOBT/FIT based screening”. In: *Evaluation and Program Planning* 62 (June 2017), pp. 73–80. ISSN: 0149-7189. DOI: 10.1016/j.evalprogplan.2017.02.007. URL: https://www.sciencedirect.com/science/article/pii/S0149718916303421 (visited on 11/21/2023).

[98] Amol Tandel et al. “Multi-document text summarization - a survey”. In: *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*. Mar. 2016, pp. 331–334. DOI: 10.1109/SAPIENCE.2016.7684115. URL: https://ieeexplore.ieee.org/document/7684115 (visited on 11/24/2023).

[99] Florence K. L. Tangka et al. “Costs of promoting cancer screening: Evidence from CDC’s Colorectal Cancer Control Program (CRCCP)”. In: *Evaluation and Program Planning* 62 (June 2017), pp. 67–72. ISSN: 0149-7189. DOI: 10.1016/j.evalprogplan.2016.12.008. URL: https://www.sciencedirect.com/science/article/pii/S0149718916302890 (visited on 11/21/2023).

[100] Oguzhan Tas and Farzad Kiyani. "A SURVEY AUTOMATIC TEXT SUMMARIZATION". scheme="ISO639-1". In: *PressAcademia Procedia* 5.1 (June 2017), pp. 205–213. ISSN: 2459-0762. DOI: 10.17261/Pressacademia.2017.591. URL: https://dergipark.org.tr/en/pub/pap/issue/33433/371658 (visited on 04/06/2024).

[101] Samia Tasnim, Md Mahbub Hossain, and Hoimonty Mazumder. "Impact of Rumors and Misinformation on COVID-19 in Social Media". English. In: *Journal of Preventive Medicine and Public Health* 53.3 (Apr. 2020), pp. 171–174. ISSN: 1975-8375, 2233-4521. DOI: 10.3961/jpmph.20.094. URL: http://jpmph.org/journal/view.php?doi=10.3961/jpmph.20.094 (visited on 02/07/2024).

[102] Shubo Tian et al. "Opportunities and challenges for ChatGPT and large language models in biomedicine and health". en. In: *Briefings in Bioinformatics* 25.1 (Nov. 2023), bbad493. ISSN: 1467-5463, 1477-4054. DOI: 10.1093/bib/bbad493. URL: https://academic.oup.com/bib/article/doi/10.1093/bib/bbad493/7505071 (visited on 06/13/2024).

[103] Robert Tibshirani. "Regression Shrinkage and Selection Via the Lasso". en. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (Jan. 1996), pp. 267–288. ISSN: 0035-9246, 2517-6161. DOI: 10.1111/j.2517-6161.1996.tb02080.x. URL: https://rss.onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1996.tb02080.x (visited on 03/06/2024).

[104] Allison Tong, Peter Sainsbury, and Jonathan Craig. "Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups". eng. In: *International Journal for Quality in Health Care: Journal of the International Society for Quality in Health Care* 19.6 (Dec. 2007), pp. 349–357. ISSN: 1353-4505. DOI: 10.1093/intqhc/mzm042.

[105] Bach Xuan Tran et al. "Coverage of Health Information by Different Sources in Communities: Implication for COVID-19 Epidemic Response". In: *International Journal of Environmental Research and Public Health* 17.10 (May 2020), p. 3577. ISSN: 1661-7827. DOI: 10.3390/ijerph17103577. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7277747/ (visited on 02/07/2024).

[106] Kate E. Trout and Li-Wu Chen. "Geographic distribution of bed occupancy during the COVID-19 epidemic in the United States: A nationwide study". In: *Health Science Reports* 4.2 (June 2021). tex.pmcid: PMC8200584, e315. ISSN: 2398-8835. DOI: 10.1002/hsr2.315. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8200584/ (visited on 09/24/2023).

[107] Iulia Turc et al. “Well-Read Students Learn Better: On the Importance of Pre-training Compact Models”. In: (2019). DOI: 10.48550/ARXIV.1908.08962. URL: https://arxiv.org/abs/1908.08962 (visited on 04/09/2024).

[108] Mojtaba Vaismoradi, Hannele Turunen, and Terese Bondas. “Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study”. en. In: *Nursing & Health Sciences* 15.3 (Sept. 2013), pp. 398–405. ISSN: 1441-0745, 1442-2018. DOI: 10.1111/nhs.12048. URL: https://onlinelibrary.wiley.com/doi/10.1111/nhs.12048 (visited on 04/07/2024).

[109] Jing Wang et al. “Systematic Evaluation of Research Progress on Natural Language Processing in Medicine Over the Past 20 Years: Bibliometric Study on PubMed”. In: *Journal of Medical Internet Research* 22.1 (Jan. 2020), e16816. ISSN: 1439-4456. DOI: 10.2196/16816. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7005695/ (visited on 11/14/2023).

[110] Mengqian Wang et al. “A systematic review of automatic text summarization for biomedical literature and EHRs”. en. In: *Journal of the American Medical Informatics Association* 28.10 (Sept. 2021), pp. 2287–2297. ISSN: 1527-974X. DOI: 10.1093/jamia/ocab143. URL: https://academic.oup.com/jamia/article/28/10/2287/6335523 (visited on 03/16/2024).

[111] Shuai Wang et al. “Integrating Extractive and Abstractive Models for Long Text Summarization”. In: *2017 IEEE International Congress on Big Data (BigData Congress)*. June 2017, pp. 305–312. DOI: 10.1109/BigDataCongress.2017.46. URL: https://ieeexplore.ieee.org/abstract/document/8029339/ (visited on 04/08/2024).

[112] Yuan Wang and Yonghao Chen. “Characterizing discourses about COVID-19 vaccines on Twitter: a topic modeling and sentiment analysis approach”. en. In: *Journal of Communication in Healthcare* 16.1 (Jan. 2023), pp. 103–112. ISSN: 1753-8068, 1753-8076. DOI: 10.1080/17538068.2022.2054196. URL: https://www.tandfonline.com/doi/full/10.1080/17538068.2022.2054196 (visited on 09/23/2023).

[113] Yuqing Wang, Yun Zhao, and Linda Petzold. “Are Large Language Models Ready for Healthcare? A Comparative Study on Clinical Language Understanding”. In: (2023). DOI: 10.48550/ARXIV.2304.05368. URL: https://arxiv.org/abs/2304.05368 (visited on 06/19/2024).

[114] David E. Wennberg et al. “A Randomized Trial of a Telephone Care-Management Strategy”. en. In: *New England Journal of Medicine* 363.13 (Sept. 2010), pp. 1245–1255. ISSN:

0028-4793, 1533-4406. DOI: 10.1056/NEJMsa0902321. URL: http://www.nejm.org/doi/abs/10.1056/NEJMsa0902321 (visited on 09/21/2023).

[115] WHO. *Cities and Public Health Crises.* Report of the International Consultation WHO/HSE/IHR/LYON/2
Lyon, Fance: World Health Organization, Oct. 2008. URL: https://iris.who.int/bitstream/handle/10665/70465/WHO_HSE_IHR_LYON_2009.5_eng.pdf (visited on 02/20/2024).

[116] WHO. *Communicating risk in public health emergencies: a WHO guideline for emergency risk communication (ERC) policy and practice.* en. Geneva: World Health Organization, 2017. ISBN: 9789241550208. URL: https://iris.who.int/handle/10665/259807 (visited on 09/23/2023).

[117] Jukkrit Wungrath and Nattapong Autorn. "Effectiveness of Line application and telephone-based counseling to improve medication adherence: A randomized control trial study among uncontrolled type 2 diabetes patients". en. In: *Health Promotion Perspectives* 11.4 (2021), p. 438. DOI: 10.34172/hpp.2021.55. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8767072/ (visited on 09/21/2023).

[118] Xi Yang et al. *GatorTron: A Large Clinical Language Model to Unlock Patient Information from Unstructured Electronic Health Records.* arXiv:2203.03540 [cs]. Dec. 2022. URL: http://arxiv.org/abs/2203.03540 (visited on 06/24/2024).

[119] Junmei Zhong and William Li. *Predicting Customer Call Intent by Analyzing Phone Call Transcripts based on CNN for Multi-Class Classification.* arXiv:1907.03715 [cs, stat]. July 2019. DOI: 10.48550/arXiv.1907.03715. URL: http://arxiv.org/abs/1907.03715 (visited on 09/21/2023).

[120] Junmei Zhong and William Li. "Predicting Customer Churn in the Telecommunication Industry by Analyzing Phone Call Transcripts with Convolutional Neural Networks". In: *Proceedings of the 2019 3rd International Conference on Innovation in Artificial Intelligence.* ICIAI '19. New York, NY, USA: Association for Computing Machinery, Mar. 2019, pp. 55–59. ISBN: 9781450361286. DOI: 10.1145/3319921.3319937. URL: https://doi.org/10.1145/3319921.3319937 (visited on 09/21/2023).

CHAPTER

7

APPENDIX FOR CHAPTER 3

7.1 Plot of topic proportions over time

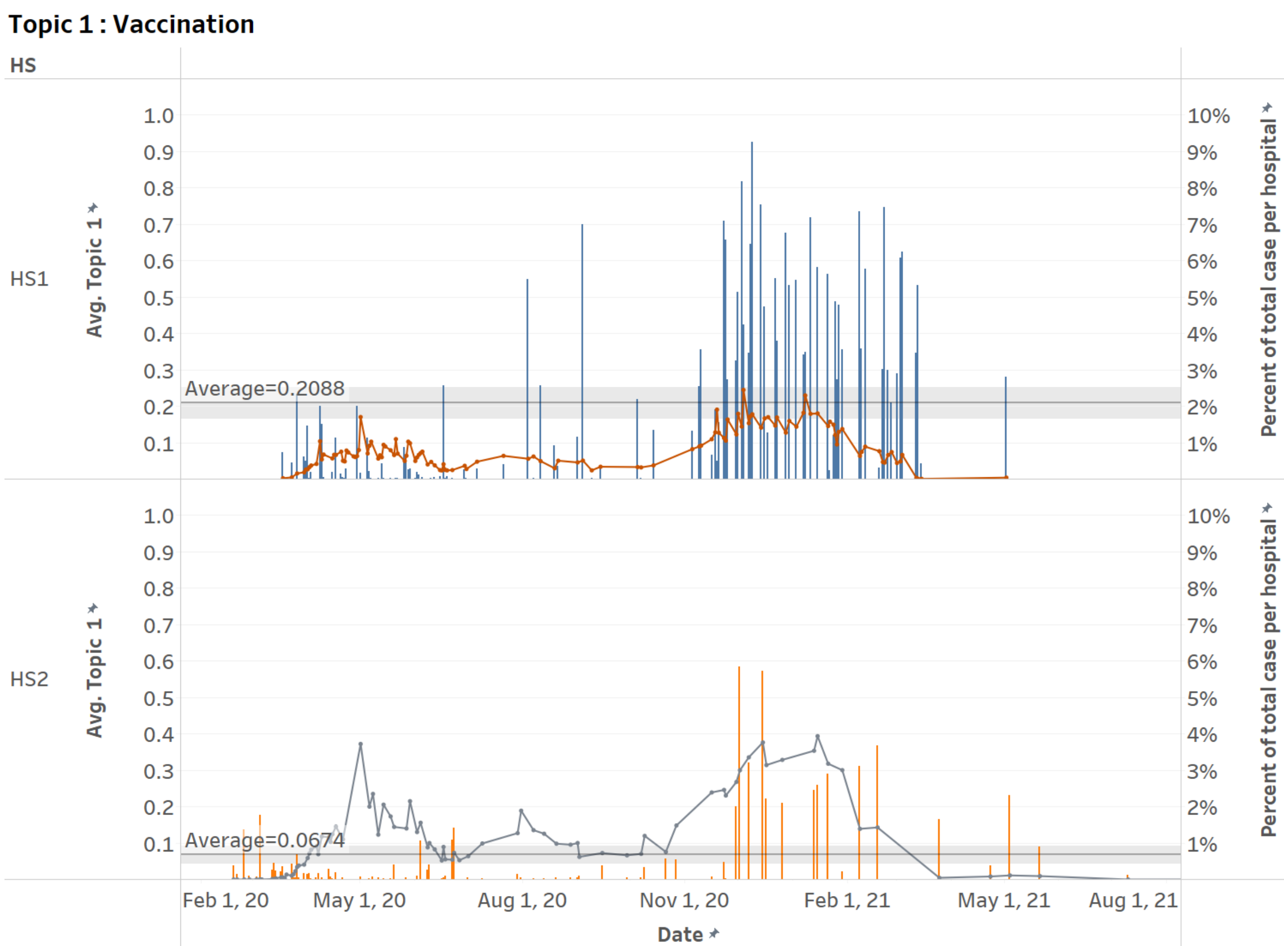


Figure 7.1: *Topic 1 Vaccination* proportion and percent of total cases per hospital across the timeline.

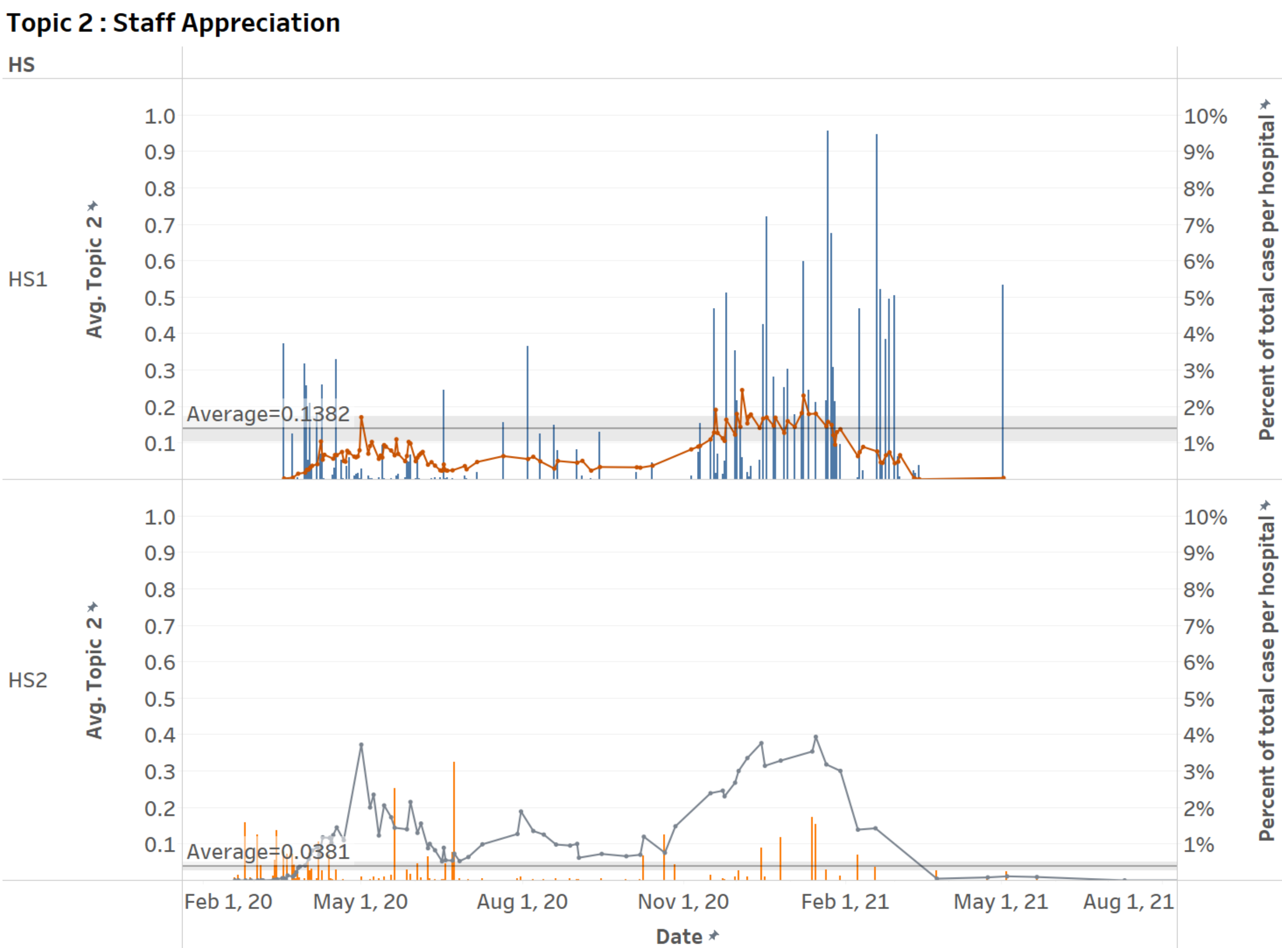


Figure 7.2: *Topic 2 Staff Appreciation* proportion and percent of total cases per hospital across the timeline.

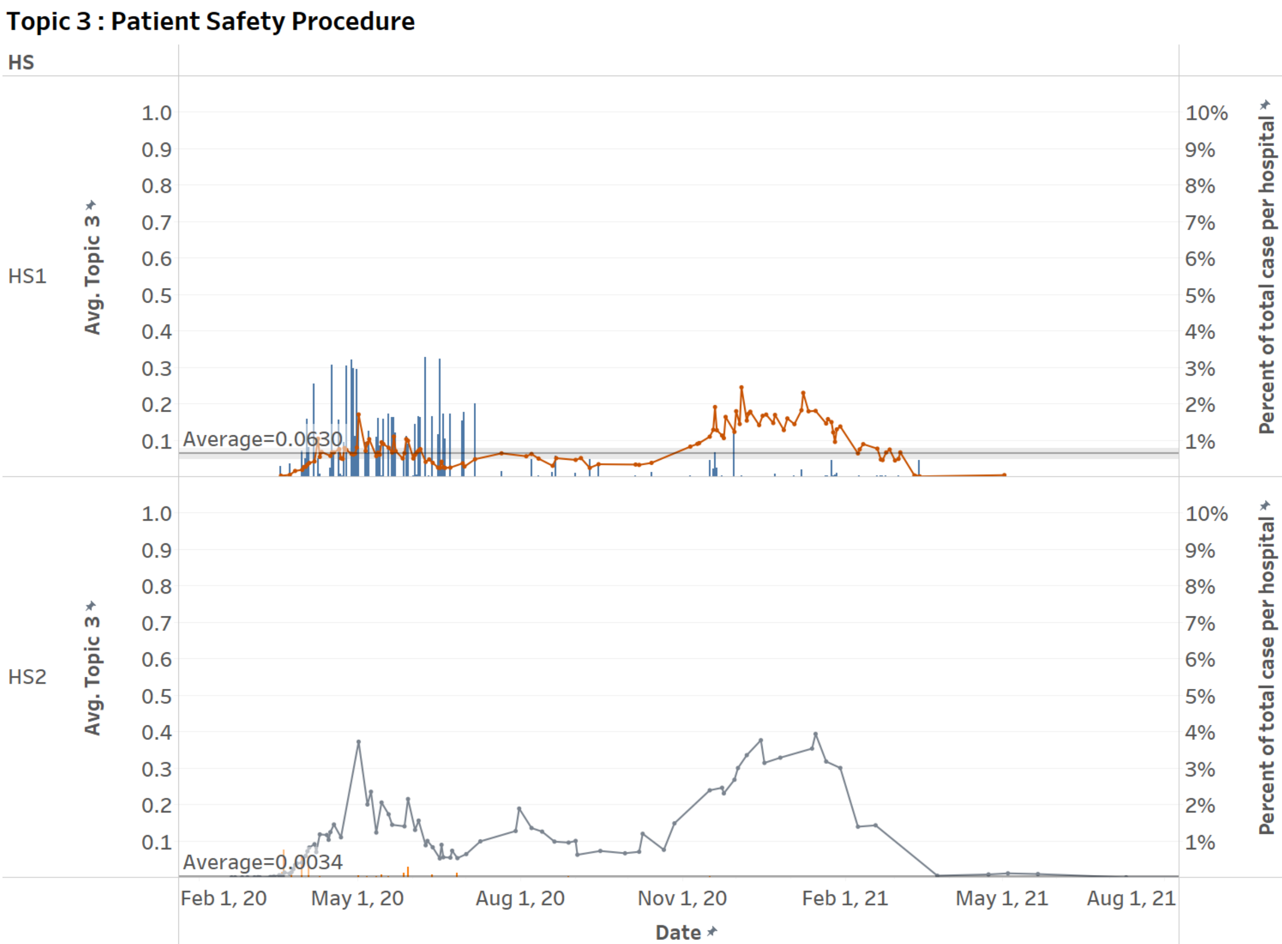


Figure 7.3: *Topic 3 Patient Safety Procedure* proportion and percent of total cases per hospital across the timeline.

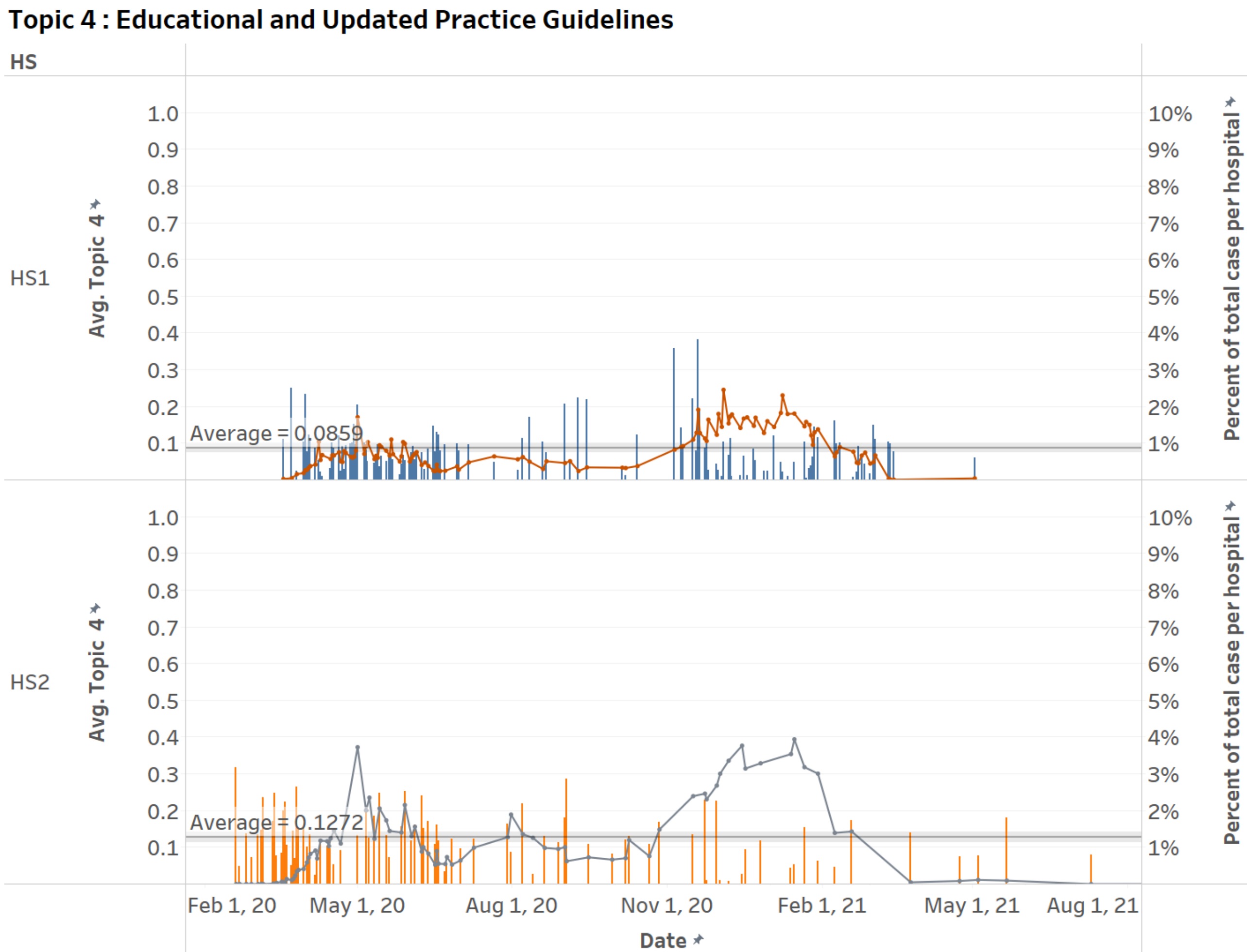


Figure 7.4: *Topic 4 Educational and Updates Practice Guidelines* proportion and percent of total cases per hospital across the timeline.

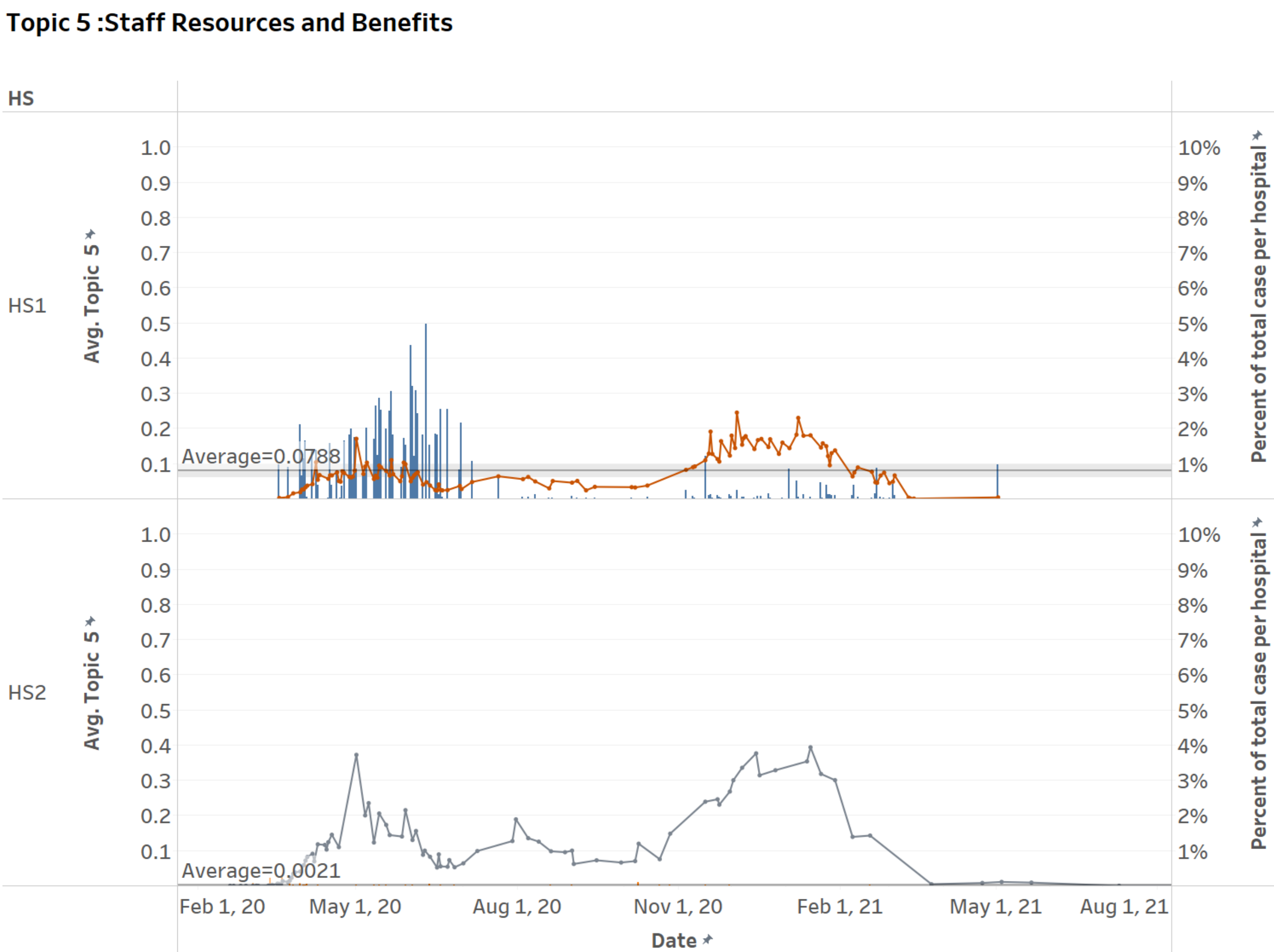


Figure 7.5: *Topic 5 Staff Resources and Benefits* proportion and percent of total cases per hospital across the timeline.

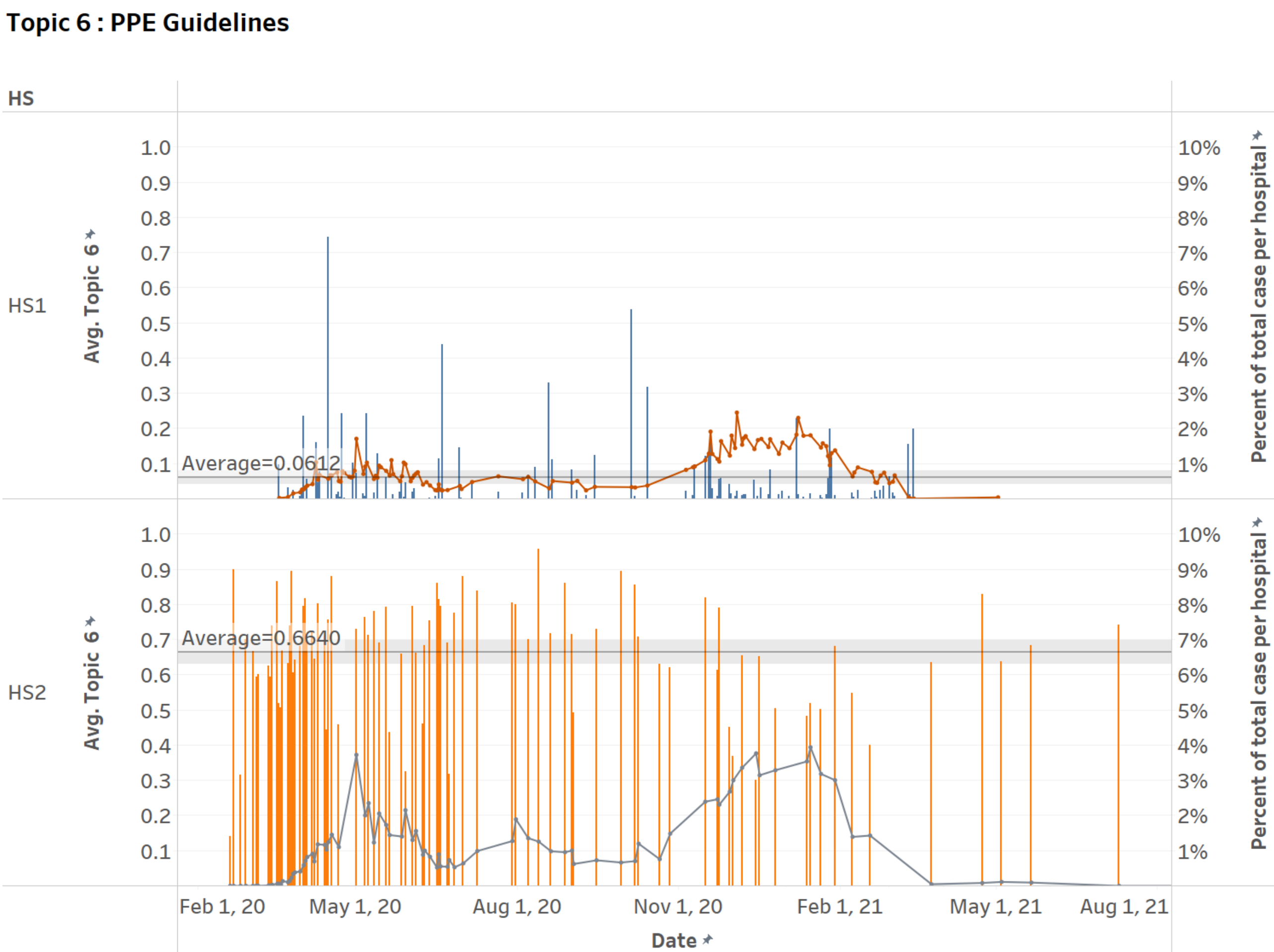


Figure 7.6: *Topic 6 PPE Guidelines* proportion and percent of total cases per hospital across the timeline.

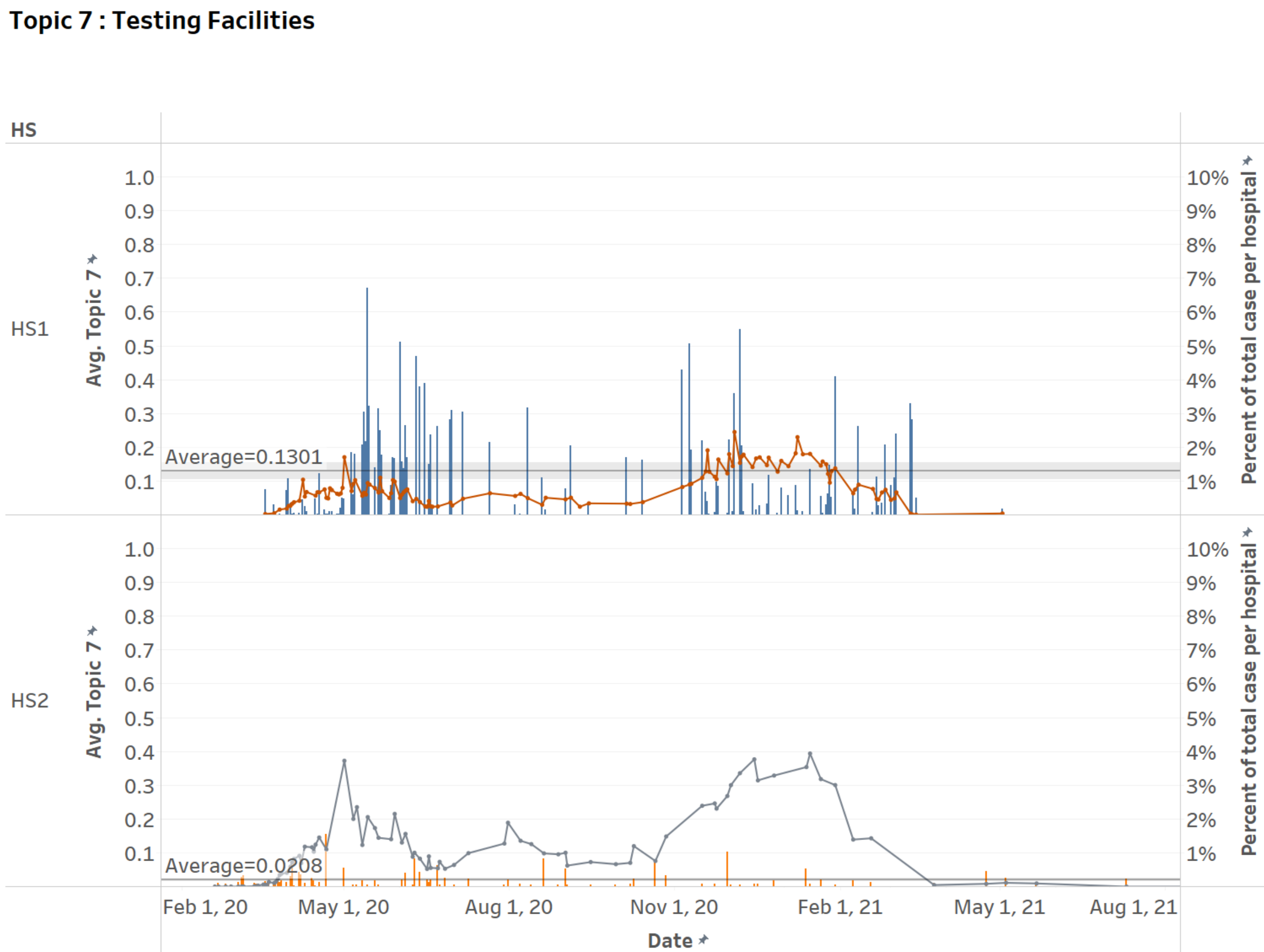


Figure 7.7: *Topic 7 Testing Facilities* proportion and percent of total cases per hospital across the timeline.

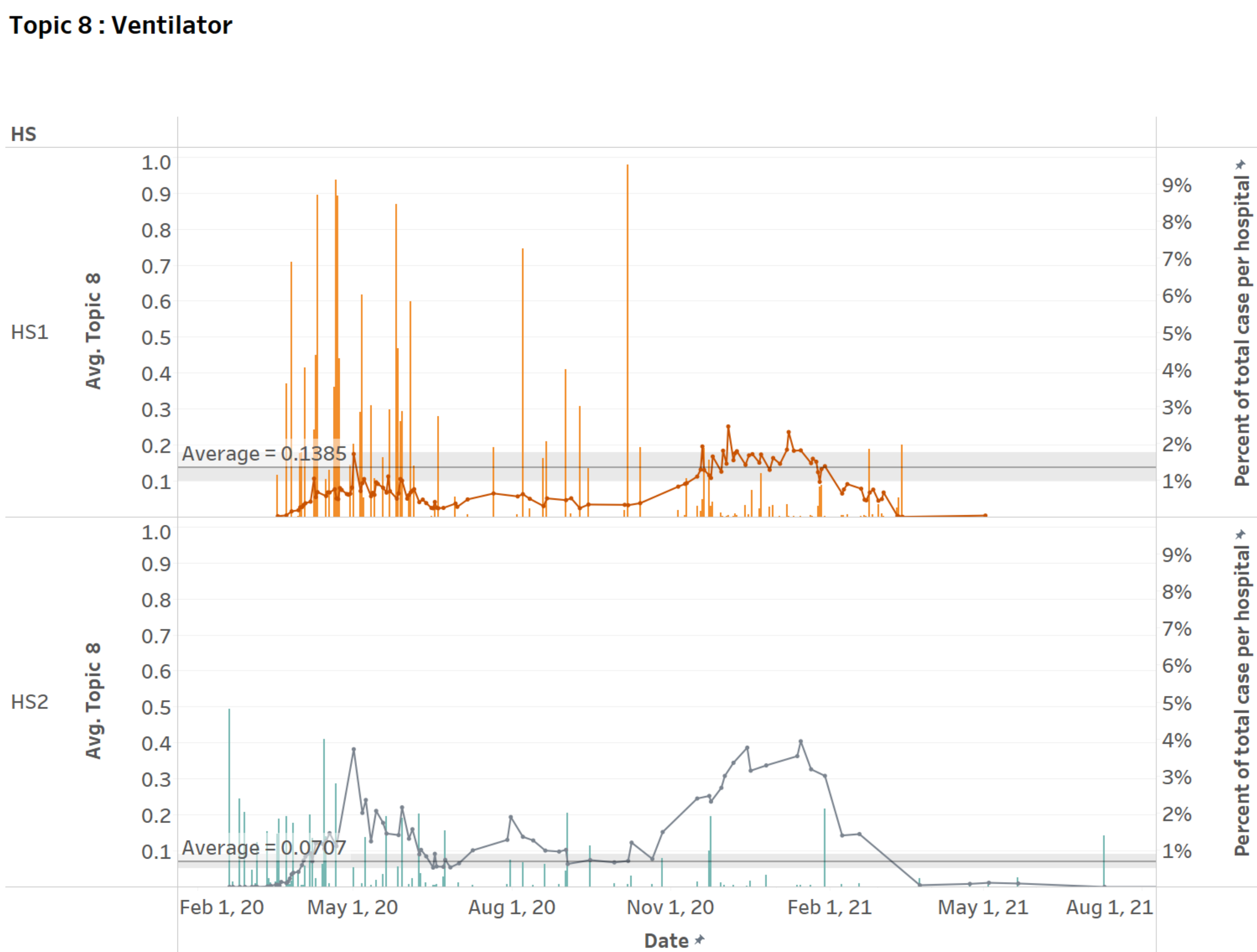


Figure 7.8: *Topic 8 Ventilators* proportion and percent of total cases per hospital across the timeline.

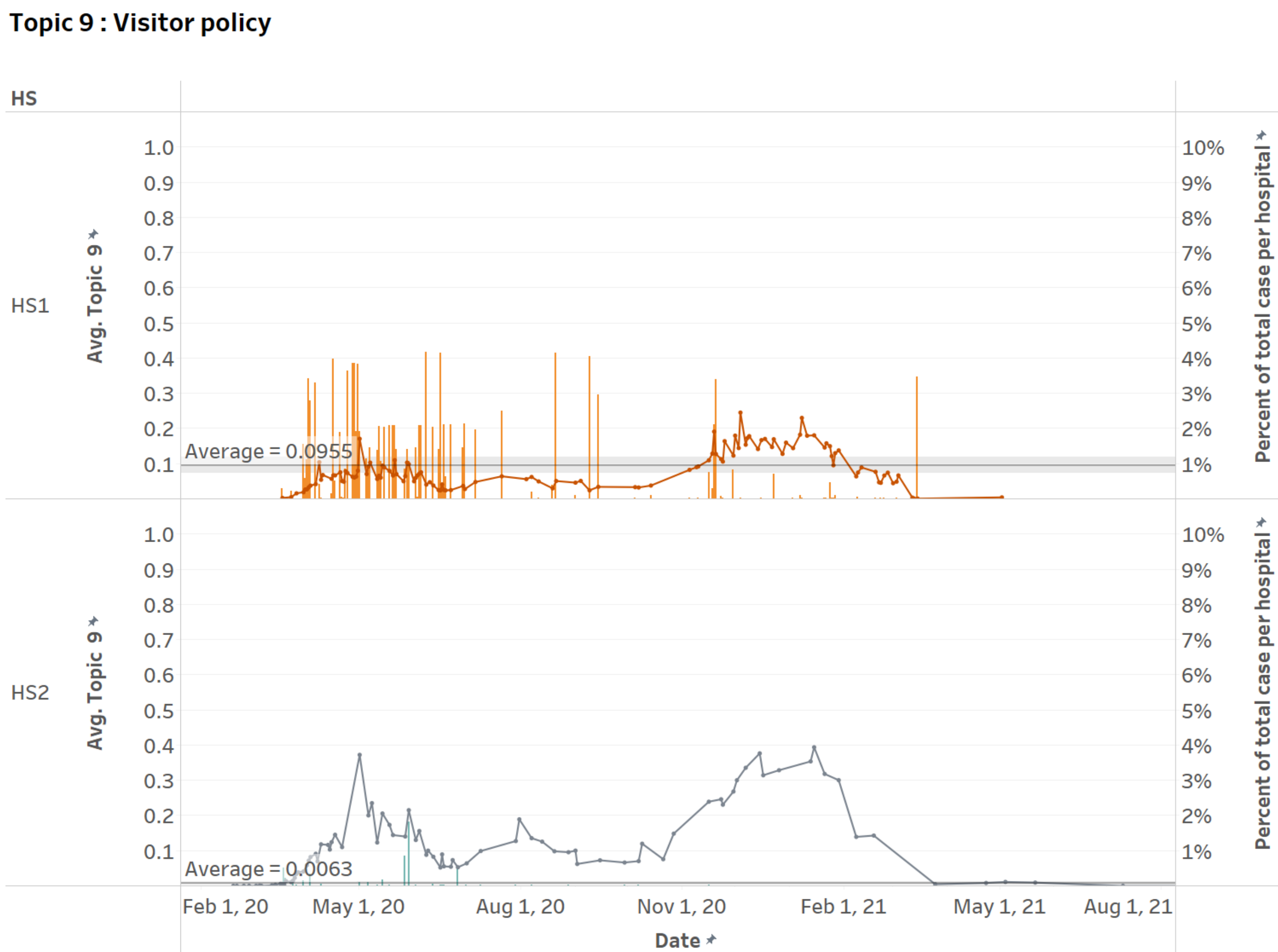


Figure 7.9: *Topic 9 Visitor Policy* proportion and percent of total cases per hospital across the timeline.

CHAPTER

8

APPENDIX FOR CHAPTER 4

8.1 Odds Ratio of Metadata Model

Table 8.1 shows the coefficient and odds ratio of all variables and its interaction in metadata model.

Table 8.1: The coefficients and odds ratios within the metadata model.

Variables	Coefficient	Odds ratio
Transcript length	1.274	3.576
Transcript length : Minimum of Customer Sentiment	0.365	1.440
Minimum of Customer Sentiment	0.249	1.283
(Intercept)	0.173	1.189
Interval	0.087	1.090
Interval : Minimum of Customer Sentiment	0.033	1.034
Interval : Number of calls	0.025	1.025
Number of calls : Minimum of Customer Sentiment	-0.013	0.987
Number of calls : Maximum of Agent Sentiment	-0.028	0.972
Interval : Maximum of Agent Sentiment	-0.054	0.947
Minimum of Customer Sentiment : Maximum of Agent Sentiment	-0.083	0.920

Table 8.1 (continued).

Variables	Coefficient	Odds ratio
Transcript length : Interval	-0.094	0.910
Transcript length : Maximum of Agent Sentiment	-0.103	0.903
Transcript length : Number of calls	-0.126	0.882
Maximum of Agent Sentiment	-0.169	0.845
Number of calls	-0.459	0.632

8.2 Odds Ratio of 4-Topic Model

Table 8.2 shows the coefficient and odds ratio of all variables and its interaction in metadata model.

Table 8.2: The coefficients and odds ratios of 4-Topic Model.

Variables	Coefficient	Odds ratio
Transcript length	2.113	8.271
Topic 4.3	1.663	5.277
Topic 4.2	0.547	1.728
Topic 4.2 : Topic 4.3	0.332	1.394
Topic 4.3 : Topic 4.4	0.331	1.393
Transcript length : Minimum of Customer Sentiment	0.323	1.382
Minimum of Customer Sentiment	0.272	1.312
Topic 4.1 : Maximum of Agent Sentiment	0.183	1.2
Topic 4.4 : Transcript length	0.139	1.149
Topic 4.1 : Interval	0.08	1.083
Topic 4.1 : Transcript length	0.079	1.083
Topic 4.2 : Number of calls	0.06	1.062
Interval	0.047	1.048
Interval : Number of calls	0.044	1.045
Minimum of Customer Sentiment : Interval	0.025	1.025
Topic 4.3 : Minimum of Customer Sentiment	0.007	1.007
Topic 4.2 : Interval	0.006	1.006
Minimum of Customer Sentiment : Number of calls	0	1
Topic 4.3 : Maximum of Agent Sentiment	0	1
Topic 4.3 : Transcript length	0	1

Table 8.2 (continued).

Variables	Coefficient	Odds ratio
Topic 4.3 : Interval	0	1
Topic 4.4 : Minimum of Customer Sentiment	0	1
Topic 4.4 : Number of calls	0	1
Topic 4	0	1
Maximum of Agent Sentiment : Number of calls	0	1
Transcript length : Interval	-0.013	0.987
Topic 4.2 : Maximum of Agent Sentiment	-0.035	0.965
Minimum of Customer Sentiment : Maximum of Agent Sentiment	-0.05	0.951
Topic 4.4 : Interval	-0.064	0.938
Topic 4.1 : Minimum of Customer Sentiment	-0.067	0.936
Maximum of Agent Sentiment : Interval	-0.076	0.927
(Intercept)	-0.078	0.925
Topic 4.1 : Number of calls	-0.15	0.861
Maximum of Agent Sentiment	-0.151	0.86
Transcript length : Maximum of Agent Sentiment	-0.162	0.851
Topic 4.1 : Topic 4.2	-0.191	0.826
Topic 4.4 : Maximum of Agent Sentiment	-0.21	0.81
Topic 4.2 : Minimum of Customer Sentiment	-0.217	0.805
Topic 4.2 : Topic 4.4	-0.222	0.801
Topic 4.3 : Number of calls	-0.239	0.788
Topic 4.1 : Topic 4.3	-0.276	0.759
Transcript length : Number of calls	-0.315	0.729
Topic 4.2 : Transcript length	-0.433	0.649
Number of calls	-0.446	0.64
Topic 4.1 : Topic 4.4	-0.991	0.371
Topic 4.1	-1.139	0.32

8.3 Odds Ratio of 5-Topic Model

Table 8.3 shows the coefficient and odds ratio of all variables and its interaction in metadata model.

Table 8.3: The coefficients and odds ratios of 5-Topic Model

Variables	Coefficient	Odds ratio
Topic 5.5	5.031	153.085
Topic 5.1 : Topic 5.5	1.778	5.915
Transcript length	1.397	4.045
Topic 5.4 : Topic 5.5	1.29	3.634
Topic 5.3 : Topic 5.5	0.927	2.528
Topic 5.2 : Topic 5.5	0.769	2.158
Topic 5.2	0.668	1.95
Topic 5.3 : Transcript length	0.477	1.612
Topic 5.5 : Transcript length	0.262	1.299
Minimum of Customer Sentiment	0.19	1.21
Topic 5.1 : Transcript length	0.179	1.196
Transcript length : Minimum of Customer Sentiment	0.139	1.149
Topic 5.3 : Number of calls	0.11	1.116
Topic 5.5 : Minimum of Customer Sentiment	0.102	1.108
Interval	0.075	1.078
Topic 5.2 : Number of calls	0.036	1.037
Topic 5.4 : Minimum of Customer Sentiment	0.035	1.036
Interval : Minimum of Customer Sentiment	0.034	1.035
Minimum of Customer Sentiment : Maximum of Agent Sentiment	0.034	1.034
Topic 5.4 : Maximum of Agent Sentiment	0.024	1.024
Topic 5.5 : Maximum of Agent Sentiment	0.02	1.021
Topic 5.2 : Interval	0.007	1.007
Topic 5.1	0	1
Topic 5.3	0	1
Topic 5.2 : Maximum of Agent Sentiment	0	1
Topic 5.4 : Transcript length	0	1
Topic 5.4 : Interval	0	1
Topic 5.1 : Number of calls	0	1
Topic 5.1 : Minimum of Customer Sentiment	0	1
Topic 5.1 : Maximum of Agent Sentiment	0	1
Topic 5.5 : Interval	0	1
Topic 5.5 : Number of calls	0	1
Transcript length : Interval	0	1

Table 8.3 (continued).

Variables	Coefficient	Odds ratio
Topic 5.1 : Interval	0	1
Topic 5.3 : Maximum of Agent Sentiment	-0.008	0.992
Topic 5.4 : Number of calls	-0.013	0.987
Transcript length : Maximum of Agent Sentiment	-0.023	0.977
Number of calls : Maximum of Agent Sentiment	-0.027	0.973
Interval : Number of calls	-0.04	0.961
Topic 5.3 : Minimum of Customer Sentiment	-0.041	0.96
Maximum of Agent Sentiment	-0.05	0.951
Topic 5.3 : Interval	-0.055	0.946
Number of calls : Minimum of Customer Sentiment	-0.068	0.934
Transcript length : Number of calls	-0.069	0.934
Topic 5.4	-0.113	0.894
Topic 5.2 : Minimum of Customer Sentiment	-0.114	0.892
(Intercept)	-0.122	0.885
Topic 5.2 : Transcript length	-0.123	0.884
Interval : Maximum of Agent Sentiment	-0.14	0.87
Topic 5.2 : Topic 5.1	-0.259	0.772
Topic 5.2 : Topic 5.4	-0.431	0.65
Number of calls	-0.521	0.594
Topic 5.2 : Topic 5.3	-0.635	0.53
Topic 5.4 : Topic 5.3	-0.822	0.44
Topic 5.1 : Topic 5.3	-0.898	0.407
Topic 5.4 : Topic 5.1	-1.273	0.28

8.4 Odds Ratio of 7-Topic Model

Table 8.4 shows the coefficient and odds ratio of all variables and its interaction in metadata model.

Table 8.4: The coefficients and odds ratios of 7-Topic Model

Variables	Coefficient	Odds ratio
Topic 7.5	2.611	13.615
Transcript length	2.252	9.508

Table 8.4 (continued).

Variables	Coefficient	Odds ratio
Topic 7.3	2.23	9.298
Topic 7.3 : Topic 7.5	1.321	3.747
Topic 7.6 : Transcript length	0.627	1.872
Topic 7.1	0.515	1.674
Topic 7.5 : Transcript length	0.5	1.648
Topic 7.5 : Topic 7.7	0.485	1.623
(Intercept)	0.417	1.518
Topic 7.7 : Transcript length	0.371	1.449
Topic 7.3 : Topic 7.7	0.358	1.43
Minimum of Customer Sentiment	0.273	1.314
Topic 7.3 : Topic 7.4	0.2	1.221
Transcript length : Minimum of Customer Sentiment	0.189	1.208
Topic 7.5 : Topic 7.6	0.182	1.199
Topic 7.1 : Topic 7.5	0.122	1.13
Topic 7.1 : Topic 7.3	0.109	1.116
Topic 7.6 : Number of calls	0.106	1.112
Topic 7.1 : Topic 7.4	0.101	1.107
Topic 7.3 : Number of calls	0.09	1.095
Topic 7.1 : Number of calls	0.088	1.092
Topic 7.4 : Minimum of Customer Sentiment	0.087	1.091
Interval	0.061	1.063
Topic 7.7 : Maximum of Agent Sentiment	0.06	1.062
Minimum of Customer Sentiment : Interval	0.053	1.054
Interval : Number of calls	0.049	1.05
Topic 7.3 : Interval	0.048	1.049
Topic 7.4 : Topic 7.5	0.038	1.038
Topic 7.1 : Topic 7.2	0.036	1.037
Topic 7.1 : Topic 7.7	0.033	1.033
Topic 7.6 : Minimum of Customer Sentiment	0.022	1.023
Topic 7.7 : Interval	0.019	1.019
Topic 7.5 : Maximum of Agent Sentiment	0.011	1.011
Topic 7.2 : Maximum of Agent Sentiment	0.011	1.011
Topic 7.2 : Interval	0.003	1.003

Table 8.4 (continued).

Variables	Coefficient	Odds ratio
Topic 7.2 : Number of calls	0.001	1.001
Topic 7.4	0	1
Topic 7.1 : Topic 7.6	0	1
Topic 7.1 : Interval	0	1
Topic 7.3 : Transcript length	0	1
Topic 7.3 : Maximum of Agent Sentiment	0	1
Topic 7.4 : Transcript length	0	1
Topic 7.5 : Minimum of Customer Sentiment	0	1
Topic 7.5 : Interval	0	1
Minimum of Customer Sentiment : Number of calls	0	1
Topic 7.7 : Number of calls	-0.002	0.998
Topic 7.2 : Minimum of Customer Sentiment	-0.003	0.997
Minimum of Customer Sentiment : Maximum of Agent Sentiment	-0.008	0.992
Topic 7.7 : Minimum of Customer Sentiment	-0.011	0.989
Topic 7.1 : Maximum of Agent Sentiment	-0.014	0.986
Topic 7.4 : Interval	-0.031	0.969
Topic 7.4 : Number of calls	-0.034	0.966
Topic 7.3 : Minimum of Customer Sentiment	-0.036	0.964
Topic 7.1 : Transcript length	-0.055	0.946
Transcript length : Interval	-0.056	0.946
Maximum of Agent Sentiment : Number of calls	-0.083	0.921
Topic 7.4 : Maximum of Agent Sentiment	-0.084	0.92
Topic 7.1 : Minimum of Customer Sentiment	-0.102	0.903
Topic 7.6 : Interval	-0.106	0.9
Maximum of Agent Sentiment : Interval	-0.115	0.891
Topic 7.2 : Topic 7.3	-0.118	0.888
Topic 7.7	-0.132	0.876
Topic 7.3 : Topic 7.6	-0.146	0.864
Topic 7.2 : Topic 7.5	-0.15	0.861
Transcript length : Number of calls	-0.192	0.826
Topic 7.6 : Maximum of Agent Sentiment	-0.2	0.819
Topic 7.5 : Number of calls	-0.205	0.814
Transcript length : Maximum of Agent Sentiment	-0.225	0.798

Table 8.4 (continued).

Variables	Coefficient	Odds ratio
Maximum of Agent Sentiment	-0.226	0.798
Topic 7.2 : Topic 7.4	-0.242	0.785
Topic 7.2 : Transcript length	-0.311	0.733
Topic 7.4 : Topic 7.7	-0.335	0.715
Topic 7.2 : Topic 7.7	-0.409	0.664
Number of calls	-0.472	0.624
Topic 7.4 : Topic 7.6	-0.636	0.529
Topic 7.2 : Topic 7.6	-0.663	0.515
Topic 7.2	-0.788	0.455
Topic 7.6	-1.005	0.366
Topic 7.6 : Topic 7.7	-1.032	0.356

8.5 Odds Ratio of 8-Topic Model

Table 8.5 shows the coefficient and odds ratio of all variables and its interaction in metadata model.

Table 8.5: The coefficients and odds ratios of 8-Topic Model

Variables	Coefficient	Odds ratio
Topic 8.6	3.972	53.112
Transcript length	2.245	9.441
Topic 8.6 : Transcript length	1.429	4.174
Topic 8.4 : Topic 8.6	0.787	2.197
Topic 8.1 : Topic 8.6	0.757	2.132
Topic 8.6 : Topic 8.7	0.712	2.038
Topic 8.2 : Transcript length	0.524	1.689
Topic 8.2 : Topic 8.6	0.504	1.655
Topic 8.7	0.462	1.588
Topic 8.6 : Topic 8.8	0.368	1.445
Topic 8.3 : Topic 8.6	0.346	1.414
(Intercept)	0.333	1.395
Topic 8.4 : Transcript length	0.284	1.328
Topic 8.6 : Maximum of Agent Sentiment	0.204	1.227

Table 8.5 (continued).

Variables	Coefficient	Odds ratio
Topic 8.1 : Topic 8.2	0.186	1.204
Topic 8.8	0.177	1.194
Minimum of Customer Sentiment	0.17	1.186
Transcript length : Minimum of Customer Sentiment	0.135	1.144
Topic 8.7 : Transcript length	0.121	1.129
Topic 8.1 : Transcript length	0.118	1.125
Topic 8.7 : Number of calls	0.117	1.125
Topic 8.6 : Minimum of Customer Sentiment	0.107	1.113
Topic 8.3 : Minimum of Customer Sentiment	0.106	1.112
Topic 8.8 : Maximum of Agent Sentiment	0.1	1.105
Interval	0.099	1.104
Minimum of Customer Sentiment : Interval	0.059	1.061
Topic 8.6 : Interval	0.057	1.059
Topic 8.3 : Interval	0.052	1.053
Topic 8.8 : Interval	0.043	1.044
Interval : Number of calls	0.041	1.042
Topic 8.1 : Maximum of Agent Sentiment	0.039	1.04
Topic 8.1 : Interval	0.032	1.033
Topic 8.2 : Number of calls	0.014	1.014
Minimum of Customer Sentiment : Maximum of Agent Sentiment	0.014	1.014
Topic 8.8 : Number of calls	0.011	1.011
Topic 8.4 : Minimum of Customer Sentiment	0.007	1.007
Topic 8.5 : Maximum of Agent Sentiment	0.001	1.001
Topic 8.4	0	1
Topic 8.5	0	1
Topic 8.1 : Topic 8.7	0	1
Topic 8.2 : Minimum of Customer Sentiment	0	1
Topic 8.3 : Maximum of Agent Sentiment	0	1
Topic 8.5 : Topic 8.6	0	1
Topic 8.5 : Number of calls	0	1
Topic 8.7 : Interval	0	1
Topic 8.8 : Transcript length	0	1
Transcript length : Interval	0	1

Table 8.5 (continued).

Variables	Coefficient	Odds ratio
Topic 8.5 : Transcript length	-0.001	0.999
Topic 8.4 : Interval	-0.005	0.995
Topic 8.5 : Interval	-0.024	0.976
Topic 8.4 : Topic 8.7	-0.027	0.973
Maximum of Agent Sentiment : Number of calls	-0.027	0.973
Topic 8.5 : Minimum of Customer Sentiment	-0.029	0.972
Topic 8.5 : Topic 8.8	-0.029	0.971
Topic 8.5 : Topic 8.7	-0.03	0.97
Topic 8.2 : Maximum of Agent Sentiment	-0.032	0.968
Topic 8.7 : Maximum of Agent Sentiment	-0.033	0.968
Maximum of Agent Sentiment : Interval	-0.034	0.966
Topic 8.3 : Number of calls	-0.035	0.966
Minimum of Customer Sentiment : Number of calls	-0.042	0.959
Topic 8.4 : Number of calls	-0.044	0.957
Topic 8.7 : Minimum of Customer Sentiment	-0.046	0.955
Topic 8.2 : Interval	-0.048	0.953
Topic 8.1 : Minimum of Customer Sentiment	-0.063	0.939
Topic 8.1 : Topic 8.5	-0.066	0.937
Topic 8.4 : Maximum of Agent Sentiment	-0.068	0.934
Topic 8.1	-0.086	0.918
Transcript length : Maximum of Agent Sentiment	-0.087	0.916
Topic 8.1 : Topic 8.8	-0.087	0.916
Topic 8.3 : Transcript length	-0.089	0.915
Topic 8.8 : Minimum of Customer Sentiment	-0.09	0.914
Topic 8.1 : Topic 8.4	-0.092	0.912
Topic 8.7 : Topic 8.8	-0.103	0.902
Topic 8.3 : Topic 8.5	-0.112	0.894
Topic 8.1 : Topic 8.3	-0.113	0.893
Topic 8.3 : Topic 8.8	-0.118	0.889
Topic 8.4 : Topic 8.5	-0.138	0.871
Maximum of Agent Sentiment	-0.149	0.862
Topic 8.3	-0.156	0.855
Transcript length : Number of calls	-0.195	0.823

Table 8.5 (continued).

Variables	Coefficient	Odds ratio
Topic 8.3 : Topic 8.7	-0.203	0.816
Topic 8.1 : Number of calls	-0.207	0.813
Topic 8.3 : Topic 8.4	-0.242	0.785
Topic 8.4 : Topic 8.8	-0.25	0.779
Topic 8.2 : Topic 8.5	-0.252	0.777
Topic 8.2 : Topic 8.7	-0.304	0.738
Topic 8.6 : Number of calls	-0.335	0.716
Topic 8.2 : Topic 8.8	-0.366	0.694
Topic 8.2 : Topic 8.4	-0.496	0.609
Topic 8.2	-0.542	0.582
Topic 8.2 : Topic 8.3	-0.609	0.544
Number of calls	-0.644	0.525

8.6 Agent's script

Following is the snapshot of the agent's script.

Voicemail: Hi, this is (Your Name), from [redacted], a Care Coordination service that partners with (Dr. XYZ.) I am calling to share a resource Dr XYZ would like to offer you to help with your vision care in between appointments. Please call me back at (Phone Number) Again, my name is (Your Name) and you can reach me at (Phone Number). I look forward to hearing from you!

QUALIFICATION CHECK BEFORE CALL

Must have:

- Medicare or Medicare Replacement Plan
- A qualifying condition and level of severity that falls under the provider's preferences
- Active management provided by the participating physician
- Fulfilled a visit with their eye care provider in the previous year
- If Glaucoma diagnosis, patient must be on drops

Nice to have:

- Supplemental plan that will pick up out of pocket cost (Tricare does not)
- Multiple drops/medications to manage

SECTION I - INTRODUCTION

IMPORTANT: Follow the appropriate process for inbound and outbound Identity Verification

Hi, may I please speak with (Patient First and Last Name)?

[Patient responds]

Hi, (patient's name). This is (Your Name) from Lumata Health, a Care Coordination service that partners with (Dr. XYZ's) calling on a recorded line. How are you today?

[respond to patient]

I'm calling you because our partner Dr. XYZ wants to provide you with a care coordinator to provide a wide range of assistance, tailored to your individual eye care needs. Dr. XYZ is committed to your health and knows how chronic health issues can make getting to appointments, medications and other activities difficult.

SECTION II - ABOUT THE PROGRAM

Your care coordinator is a certified professional who partners with Dr. XYZ to provide you extra support **over the phone** in between your appointments to help guide you towards better eye health. We are here to address questions and concerns about your treatment, provide education on (patient condition), help you monitor your vision for changes, and make sure you have the right support in place to preserve your vision.

We can be flexible on how often we talk. Our phone calls can be as frequent or infrequent as you'd like, and you can stop at any time. Care Coordination is covered by Medicare and is subject to your annual deductible and coinsurance (usually 20%), but if you have a secondary or supplemental insurance plan, your coinsurance may be covered.

- If a patient asks about insurance coverage specifics: A claim will be submitted every time we spend at least 30 minutes managing your care within a calendar month. So let's say we talk every 2 months, then you'd have around 6 claims submitted in a year for these encounters. Since you (do/don't) have secondary coverage, this program should (be no cost to you/include your typical cost sharing at 20%, which usually amounts to less than $10-12).

Figure 8.1: Agent's script page 1

- The practice will bill your insurance every time we spend at least 30 minutes providing care coordination in a calendar month.Care coordination services like this actually save medicare an average of $70 per patient per month by keeping your disease from getting worse which would require more expensive treatment.

SECTION IIA- ENROLLMENT
<u>(Enrollment Specialist)</u>: **So let's schedule time for you to work with your Care Coordinator.** What days/times are you most available for the Care Coordinator to call you? (document response in "General Availability" field) . • If patient doesn't agree ⟶ Okay, no problem. What concerns do you have about working with a Care Coordinator so I can update Dr. XYZ? • If agrees ⟶ **document response and** prepare them for the call with Coordinator Great, I have you scheduled for around ___ AM/PM on ____ (date).Your care coordinator will call you to ask a few questions related to your (Patient Condition) care and create a care plan with information personalized to you and your needs. The care plan is meant to be a helpful resource for you. It takes about 10 minutes or so to complete. We will reach out to remind you of the scheduled call. Would you like the reminder to be a text or phone call?

Figure 8.2: Agent's script page 2

CHAPTER

9

APPENDIX FOR CHAPTER 5

9.1 List of combined words

Table 9.1: List of combined words

word 1	word 2
assessment	feedback
best	practices
breast	cervical
control	program
bowel	preparation
cancer	screening
cancer	screenings
care	clinics
care	providers
chronic	disease
client	reminders
clinic	partners
clinic	champions

word 1	word 2
clinical	services
clinical	service
colonoscopy	screening
colorectal	cancer
community	guide
community	health
community	education
community	partners
contextual	factors
cost	data
cost	per
costs	per
covid-19	pandemic
crc	screening
crc	interventions
crc	screenings
crccp	grantees
crccp	awardees
crccp	funding
crccp	clinics
crccp	learning
crccp	programs
crccp	program
crccp	recipients
data	collection
data	collected
diagnostic	follow-up
ebi	implementation
ebis	implemented
economic	evaluation
early	detection
education	outreach
electronic	health
evidence	base
evidence-based	interventions

word 1	word 2
fecal	occult
federally	qualified
field	guide
fit	kits
fit	kit
fobt	screening
funding	cycle
future	research
future	studies
health	systems
health	system
health	care
health	centers
health	programs
health	departments
health	records
health	program
implement	ebis
implementation	economics
implementation	science
implementation	process
implementing	ebis
increase	screening
incremental	cost
integrated	approaches
integrated	delivery
integrated	interventions
kit	return
learning	laboratory
learning	collaborative
lessons	learned
mailed	fecal
mailed	fit
mass	media
medically	underserved

word 1	word 2
multicomponent	interventions
multiple	cancer
multiple	cancers
multiple	ebis
nonclinical	costs
occult	blood
outreach	events
partner	clinics
patient	navigation
patient	reminders
patient	reminder
people	screened
percentage	points
population	level
preventive	services
per	additional
primary	care
prior	studies
program	planning
program	activities
promotion	activities
provider	assessment
provider	reminders
provider	reminder
public	health
qualified	health
quality	improvement
quality	assurance
quality	indicators
reducing	structural
reminder	systems
reminders	provider
research	needed
resources	required
return	rates

word 1	word 2
sample	size
screening	rates
screening	promotion
screening	uptake
screening	rate
screening	programs
screening	test
screening	tests
screening	program
screening	provision
screening	services
screening	status
screening	use
screening	process
service	delivery
services	task
several	limitations
small	media
state	health
stool	test
stool	tests
stool	testing
structural	barriers
study	findings
support	staff
supporting	activities
successfully	screened
task	force
technical	assistance
training	ta
tribal	organizations
united	states